Thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Relationship between segmental speech errors and intelligibility in acquired dysarthria

Naomi Rachel Miller

October 2018

City, University of London
Department of Language and Communication Science
School of Health Sciences

Table of Contents

List of Tables

Table 4.2. Mean consonant error rates across the speakers with dysarthria, separated according to word position (initial or final).

Table 4.3. Phonetic-contrast confusions at word-initial (C1) position. The mean percentage error is a measure of the prominence of the error relative to all C1 errors, as described in the text.

Table 4.4. Phonetic-contrast confusions at word-final (C2) position.

Table 4.5. Vowel substitutions quantified in terms of the mean percentage error (MPE), a measure of the average prominence of the error with respect to all other vowel errors. The final column shows the predominant error direction and the MPE for that direction.

Table 4.6. Vowel confusions for three female speakers with different vowel accuracies (shown in brackets). The first number is the total percentage error summed over both directions. The number in brackets is the percentage error for the direction indicated by the phoneme order in the first column. For example, for the *bed* - *bid* confusion, the vowel confusions were unidirectional for S7 and S2, while S4 yielded an error rate of 7.8% for *bed* → *bid* and 1.3% for *bid* → *bed*. The top five confusions for each speaker (i.e., summed over both directions) are shaded in grey.

Table 4.7. Mean consistency score (final column) for all speakers. The word accuracy for each speaker is also shown. There is a negative correlation (Pearson's r = -0.60, p = 0.06) between these two quantities such that speakers who were more intelligible yielded errors of lower consistency.

Table 4.8. Word-frequency data (taken from the CGN) for a subset of the target words used in this study. The bold line indicates the boundary between frequency ratings of 3 and 4. The final column shows the increase in the number of occurrences relative to the next most common target word.

Table 4.9. Consonant contrast categories observed reasonably consistently in the present study. The second column discusses various aspects of each category, as explained in the text. Bold typeface denotes that the predominant error direction observed for the category is hypothesised to be genuine (i.e., not an artefact of the linguistic features of the word list).

Table 5.1. Demographic information and accuracy metrics for the control subjects assessed using orthographic transcription of the single-word reading stimuli.

Table 5.2. Semi-quantitative descriptions of the distributions of C1 error rates across the cohort of neurotypical speakers. The data are organised such that phonemes closer to the top of the table are more vulnerable, to the best of the author's judgment.

Table 5.3. Semi-quantitative descriptions of the distributions of C2 error rates across the cohort of neurotypical speakers. Phonemes closer to the top of the table were judged to be more vulnerable.

Table 5.4. Mean consonant error-rates for speakers with dysarthria, displayed in order of decreasing frequency. Error rates were derived from Table 4.2, but have been rounded up or down to the nearest 1%. In addition, errors below a certain frequency have been grouped together.

Table 5.5. Mean and median vulnerability rates for the most prominent directional contrast errors. The last column shows the result of significance testing, either using either the Student's t-test (above the bold line) or the Mann-Whitney U test (below the bold line). Within each test type, the data are presented in order of decreasing *p*-value. Grey shading means that there is *no* evidence that the contrast error occurs more often in speakers with dysarthria, while unshaded means that there is *strong* evidence that the error is dysarthric ($p < 0.05$).

Table 6.1. List of consonant contrast categories tested in the multiple-choice study along with predictions for the predominant error direction that would be observed. "None" implies that there was no clear evidence for making a prediction about directionality.

Table 6.2. Comparison of word accuracies in the free- and forced-response modes. The final column ("Diff") shows the absolute increase in accuracy between the two response modes.

Table 6.3. Fleiss' kappa for the multiple-choice study. The consistency metric used to measure inter-rater agreement in the free-response mode is shown for comparison. The final row shows the correlation between consistency and Fleiss' kappa and between MC word accuracy and Fleiss' kappa.

Table 6.4. Pearson's *r* for consonant error rankings in the two response modes for each speaker. The speakers are displayed in order of decreasing intelligibility (based on the free-response study). The *p*-values marked with an asterisk are significant assuming a Bonferroni-corrected alpha level of $(0.05 / 8) = 0.0063$.

Table 6.5. Mean vowel vulnerability rates in the forced-choice study. The data are displayed in order of decreasing vulnerability (summed over the two directions) from top to bottom. The number of occasions on which each directional confusion was tested is shown in parentheses.

Table 6.6. Pearson's *r* for the vowel error rankings in the two response modes for each speaker, where the speakers are displayed in order of decreasing intelligibility (based on the free-response study). The *p*-values marked with an asterisk are significant assuming a Bonferroni-corrected alpha level of $(0.05 / 8) = 0.0063$.

Table 6.7. Importance of consonant contrast categories in Dutch dysarthria, as judged by three criteria. The categories highlighted in grey are classed as "important". The numbers in Column 1 facilitate comparison with Fig. 6.3.

Table 6.8. Importance of vowel contrast categories in Dutch dysarthria, as judged by three criteria. The categories highlighted in grey are classed as "important". The numbers in Column 1 facilitate comparison with Fig. 6.8.

Table 7.1. Technique used to correct for guesswork. The speaker was describing his grandson's football match, and in the author's estimation, the best possible transcription of the utterance would have been *00 gewonnen heeft of verloren* ("00 has won or lost"), where the zeroes denote unintelligible syllables. For each listener, the transcribed words that were classed as "intelligible" based on the consensus method are shown in blue font.

Table 7.2. Characteristics and intelligibility scores of all the monologues assessed in the study. The final two columns show the word accuracies obtained in single-word reading (free-response and multiple-choice). Alternate speakers have been shaded differently (grey vs. white).

Table 7.3. Results of correlation analysis between spontaneous-speech intelligibility and three different characteristics of the speakers' monologues: utterance length, speech rate and dysfluency.

List of Figures

Figure 4.8. Mean error rate (± 1 standard deviation) for the five word-frequency ratings, where a rating of 1 denotes the highest lexical frequency.

Figure 4.9. Mean error rate (± 1 standard deviation) for the five word-frequency ratings, where a rating of 1 denotes words of the highest frequency.

Figure 5.1. (a) Word accuracy and (b) C1 accuracy for all speakers. The dotted pattern represents neurotypical speakers, while the filled (grey) bars represent speakers with dysarthria.

Figure 5.2. C1 contrast-error rates in (a) speakers with dysarthria and (b) control speakers. The two colours denote the two directional errors. Error rates represent the number of times that the directional error was observed divided by the total number of C1 errors made by the speaker. These values were then averaged over the whole cohort to yield the mean percentage error. Blue (orange) refers to the following directions: devoicing (voicing) for stop and fricative voicing errors (S_v and F_v); deletion (addition) for initial cluster vs. singleton (IC/sng) and /h/ vs. null (h/null) confusions; backing (fronting) for stop, nasal and fricative place errors (S_p, N_p and F_p); fricative → stop (stop → fricative) for the category F/S; fricative → /r/ (/r/→ fricative) for the category F/r; /r/ → /l/ (/l/ → /r/) for the category r/l; and nasal → stop (stop → nasal) for the category N/S.

Figure 5.3. C2 contrast-error rates in (a) dysarthric and (b) control speakers. Blue (orange) refers to: deletion (addition) for final cluster vs. singleton (FC/sng) and final consonant vs. null (C/null); backing (fronting) for nasal place errors (N_p); fricative → stop (stop → fricative) for the category F/S; fricative → /r/ (/r/→ fricative) for the category F/r; /r/ → /l/ (/l/ → /r/) for the category r/l; and nasal → stop (stop → nasal) for the category N/S.

Figure 5.4. Vowel confusions in (a) dysarthric and (b) control speakers. Blue denotes the error direction implied by reading each label from left to right. Thus, the errors are monophthong → diphthong (diphthong → monophthong), /ɛ/ → /ɪ/ (/ɪ/ →/ɛ/), /a:/ → /ɑ/ (/ɑ/ → /a:/), and so on.

Figure 5.5. Monophthong confusions for (a) dysarthric and (b) control speakers. The thickness of each arrow is proportional to the sum of the MPEs across both directions. The arrow head indicates the predominant error direction; two arrow heads are shown when the errors are bidirectional.

Figure 5.6. Word-accuracy values (a) excluding and (b) including categories that are likely to be non-dysarthric. (Dotted pattern: neurotypical speakers; filled bars: speakers with dysarthria). The blue and orange lines show 95% and 97.5% confidence levels for dysarthria detection, respectively.

Figure 6.1. Comparison of word accuracies in the free-response and multiple-choice (MC) studies. The data are presented in order of increasing word accuracy for the free-response mode.

Figure 6.2. Mean vulnerability rates for the consonant contrast categories in dysarthric speakers. Blue (orange) refers to the error direction: devoicing (voicing) for stop and fricative voicing errors (voice); backing (fronting) for stop, nasal and fricative place errors (S_p, N_p and F_p); stop → fricative (fricative → stop) for the category S/F; nasal → stop (N/S); /r/ → fricative (fricative → /r/); /r/ → /l/ (/l/ → /r/); and addition (deletion) for /h/ vs. null (h/null), initial cluster vs. singleton (IC/sng), and final cluster vs. singleton (FC/Sng).

Figure 6.3. Sum of ranked errors for the consonant contrast categories as assessed via the forced- and free-response modes. The category codes are shown in the table beneath the figure.

Figure 6.4. Relationship between the total sum of the ranks for the 26 consonant categories.

Figure 6.5. Vulnerability rates (y-axis) for individual speakers (x-axis) for eight of the consonant contrast categories. The speakers are presented in order of increasing severity from left to right (i.e., in order of decreasing word accuracy in the free-response mode). Blue (orange) shading refers to the error direction: devoicing (voicing) for the category 'voice'; backing (fronting) for fricative place; stop → fricative (fricative → stop); /r/ → fricative (fricative → /r/); addition (deletion) for final cluster vs. singleton and initial cluster vs. singleton; nasal → stop; and /r/ → /l/ (/l/ → /r/).

Figure 6.6. Monophthong confusions from Table 6.5 (forced choice). The thickness of each arrow is proportional to the mean vulnerability rate (summed over the two directions), while the arrow head indicates the predominant direction (two arrow heads denote bidirectional errors).

Figure 6.7. Monophthong confusions from the free-response study. The thickness of each arrow is proportional to the total number of occasions on which the confusion was observed, summed over both directions and over all eight speakers.

Figure 6.8. Sum of ranked errors for the directional vowel confusions as judged via the forced- and free-response modes. The confusion codes are shown in the legend beneath the figure. The last two confusions were only tested in one direction in the multiple-choice mode.

Figure 6.9. Relationship between the total sum of the ranks in the two response modes for the 24 vowel categories. Pearson's $r = 0.62$, $p = 0.006$ (one-tailed).

Figure 7.1. Example of a case where it was difficult to locate the precise start-point of the utterance. The initial phoneme could not be identified, but it was thought to be a vowel. The yellow trace represents the intensity contour. The vertical red dotted line indicates the chosen transition point.

Figure 7.2. A case where it was difficult to locate the end point of the utterance. The transition was from a schwa to silence. The vertical red dotted line indicates the chosen transition point.

Figure 7.3. Pearson's *r* and one-sided *p*-values for the relationship between SSI and word accuracy calculated from (a) the free-response mode (n = 10) and (b) the MC mode (n = 8). The red circles in the left-hand figure represent speakers who were not assessed in the forced-choice mode.

Figure 7.4. Correlation between SSI and four different accuracy metrics from the orthographic transcription of single words (see x-axis labels). The datapoints of three speakers (S4, S8 and S9) have been labelled, to facilitate the discussion in Section 7.4.1. The dotted line shows the best-fit linear regression, the slope of which is equal to Pearson's *r*.

Figure 7.5. Correlation between dysfluency (the proportion of the duration of the monologue occupied by between-utterance pauses) and SSI (r = -0.58, one-tailed p = 0.04).

Figure 8.1. Monophthong confusions involving contrasts in vowel height and backness that were shown to be important in the present cohort.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisors Jo Verhoeven and Peter Mariën for their continuous support and motivation and for sharing their immense knowledge. No obstacle seemed to faze them and data acquisition would not have been possible without the considerable effort they made to ensure that I had access to all the necessary people and materials. Jo has been a source of tremendous guidance and inspiration throughout. His ability and willingness to respond thoughtfully on any matter are truly remarkable. His deep theoretical knowledge combined with his common-sense approach have been invaluable in helping me interpret the findings and draw conclusions. I will always be grateful for Peter's contribution. His knowledge, experience and enthusiasm were crucial factors in finalising the project design and setting up the interview process. It is immensely sad that we were denied the opportunity to reflect on the results together, as I know his insights would have been fascinating. He will be greatly missed.

Secondly, I am extremely grateful to all the Speech and Language Therapists at the Ziekenhuis Netwerk Antwerpen who gave up their valuable time in helping to recruit participants, access hospital records, and carry out listening sessions: Ineke Wilssens, Nancy Hufkens, Griet van Gestel, Dorien van Den Bulck, Marleen Merckx, and Lien Van Rompaey. You all made me feel very welcome and "ik vond het heel gezellig!" Special thanks must go to Ineke who showed tremendous support and enthusiasm, and who spent a considerable amount of time not only on practical issues, but also on discussing details of the methodology and findings. Her passion and skills as a therapist are inspirational.

Thirdly, I would like to thank all the participants who contributed to the study. To the speakers, I greatly appreciate your generosity in giving up your time, and the enthusiasm with which you embraced all the speech tasks. I enjoyed meeting you all and listening to your stories. I am also extremely grateful to all the listeners, and especially to those who helped recruit others. In particular, Stefanie Keulen went to great efforts to organise the listening session at Brussels University and to recruit listeners from among her colleagues. Special thanks are also due to Carla Verweerden and Harmen Rooms, who went out of their way to find listeners from among their friends and family.

The opportunities I have had to discuss my work with colleagues at City have been immensely valuable. The feedback and guidance I received at my transfer viva from Katerina Hilari, Shula Chiat and Rachael Anne-Knight were of crucial importance in pinpointing the focus of the project. I am also grateful to the staff and students in the department who attended the annual PhD seminars and asked insightful questions. Support and practical help also came from Becky Moss, Maddie Pritchard and Ellen Thaels.

Finally, I am deeply grateful for the practical and emotional support provided by my friends and family. To Harrie, Tonnie and Cyriel van Aar: Thank you for stepping in with baby-sitting, often at short notice, and for giving me feedback on the speech stimuli and listening studies. To my friends in Eindhoven – Rotem, Anrie, Olga, and Rita: Although you didn't succeed very often, thank you for trying to drag me away from my computer and for providing much needed entertainment and distraction, as well as practical help with Kyla and Hannah. Thanks are also due to my parents, Liz and Steve, who crossed a sea to fulfil their babysitting duties and who were always available to discuss scientific matters of any kind, from ethics proposals through to study design and statistics. Above all, I owe a huge debt of gratitude to Rajan and Hannah van Aar. Rajan has supported and encouraged me every step of the way and has made numerous contributions to the project. During the early stages, he never tired of answering the question "Is this a Dutch word?" and he was my main language advisor throughout. We have had countless discussions about data interpretation and methodology, and his impeccable logic proved to be of great value. The last acknowledgement goes to my precious daughter, Hannah. She has shown patience, love and understanding in measures beyond her years and she never fails to brighten my day.

Abstract

This thesis began with the assumption that the clinical goal is to identify articulatory, segmental speech errors. The main objective was to explore the potential of using *phonetic-contrast analysis* for this purpose in Belgian Dutch speakers with dysarthria. In this approach, the speaker reads a list of monosyllabic words, and the listener either transcribes each word orthographically (*open* mode) or chooses from among the target and a set of foils (*closed* mode). The perceived phonemic substitutions are then coded as contrasts in a single phonetic feature, e.g., initial-stop voicing. The secondary objectives of the thesis were to (a) identify vulnerable phonetic contrasts in Belgian Dutch dysarthria and (b) assess the correlation between word-reading accuracy and intelligibility in spontaneous speech.

A phonemically-balanced word list was developed (117 words). It was read by 10 subjects with dysarthria (due to various aetiologies) and 8 neurotypical controls, all from the Antwerp region. The speakers with dysarthria also delivered monologues on topics of their choosing. Online listening sessions were conducted in which the single-word stimuli were identified using both an open and a closed mode. The monologues were assessed using a syllable-accuracy metric derived from orthographic transcription (Lagerberg et al., 2014).

Phonetic-contrast analysis showed significant promise with regard to consonants: more than 78% of substitutions could be coded using 13 contrast categories. Vowel confusions, however, did not typically lend themselves to categorisation based on a single phonetic feature (e.g., height), partly due to the configuration of the Antwerpian vowel space. Contrasts that were no more vulnerable in dysarthria than in the control group included the voicing of word-initial stops and confusions between high, front vowels. Word accuracy for the dysarthric group was significantly higher for the open than the closed mode (mean absolute difference 13.1% ± 6.9%). The two response modes also yielded different error profiles. In each mode, for each dysarthric speaker, vowel and consonant contrasts were separately ranked according to error rate. Pearson's r between the ranks for the two modes, calculated for each speaker, ranged from 0.34 - 0.72 for consonants and 0.17 - 0.86 for vowels. Prominent consonant confusions included initial-stop devoicing, singletons perceived as clusters (a distortion error), and confusions between fricatives and another manner of articulation. Vowel confusions typically corresponded to either (i) reductions, such as shortening or monophthongisation, or (ii) confusions between phonemes that are relatively close together in F1-F2 space in the Antwerp accent. A correlation of 0.76 (one-tailed $p < 0.01$) was observed between word intelligibility and intelligibility in spontaneous speech. Overall, the findings suggest a complex set of interactions between speaker characteristics, listener characteristics and the methods used to elicit and analyse speech.

List of Abbreviations

AAC	Augmentative and alternative communication
ALS	Amyotrophic lateral sclerosis
AMRs	Alternating motion rates
AOS	Apraxia of speech
C1	Word-initial consonant
C2	Word-final consonant
CGN	Corpus Gesproken Nederlands
CT	Computed tomography
CVA	Cerebrovascular accident
DME	Direct magnitude estimation
EAI	Equal-appearing interval
FA	Friedreich's ataxia
FDA	Frenchay Dysarthria Assessment
IPA	International Phonetic Alphabet
LME	Linear mixed effects
MC	Multiple choice
MCA	Middle cerebral artery
MMSE	Mini Mental State Examination
MPE	Mean percentage error
MRI	Magnetic resonance imaging
NSVO	Nederlandstalig Spraakverstaanbaarheidsonderzoek
PCA	Posterior cerebral artery
PCC	Percentage of correct consonants
PD	Parkinson's disease
R-BANS	Repeatable Battery for the Assessment of Neuropsychological Status
RDA	Radboud Dysarthria Assessment
RP	Received pronunciation
S	Speaker

SD	Standard deviation
SIT	Sentence Intelligibility Test
SLT	Speech and language therapy / Speech and language therapist
SSI	Spontaneous speech intelligibility
SWR	Single-word reading
V	Vowel
VAS	Visual analogue scale
VIF	Variance inflation factor
VOT	Voice onset time
WAIS-IV	Wechsler Adult Intelligence Scale IV
ZNA	Ziekenhuis Netwerk Antwerpen

1. Introduction

1.1. Acquired dysarthria

In the classification of motor speech disorders, a distinction is made between apraxia of speech (AOS) and dysarthria. AOS is considered to be impaired ability to plan or program the sensorimotor commands needed for producing speech (Duffy, 2005). Dysarthria is defined as impaired ability to execute speech, due to a disturbance in neuromuscular control (Palmer, 2005). In dysarthria, impairments result from damage to the central or peripheral nervous system, which may cause "weakness, slowing, incoordination, altered muscle tone, and inaccuracy of oral and vocal movements" (Palmer & Enderby, 2007). The effect of these impairments on speech intelligibility is highly variable, but even when intelligibility is not severely affected, dysarthria can have a significant negative impact on an individual's communicative experiences, which can affect their well-being and quality of life (Walshe & Miller, 2011). Acquired dysarthria may arise due to a neurological insult, in which case it tends to be stable, or from a neurological disease, in which case the nature and severity of the speech deficits may progress over time. The incidence of dysarthria is unknown; however, a common cause is stroke and the UK incidence of dysarthria associated with stroke alone is of the order of 30,000 cases per year (RCSLT, 2009).

1.2. Dysarthria assessment

This thesis upholds that the aims of impairment-based dysarthria assessment should be:

1) To determine the speaker's level of intelligibility
2) To identify their prototypical speech deficits
3) To determine which deficits are most detrimental to their intelligibility.

In this thesis, intelligibility is defined as the degree to which the intended speech signal is recovered by the listener (Kent et al., 1989), with any reduction in intelligibility considered to be confined to the speech *production* phase. Accordingly, intelligibility is an indicator of dysarthria severity. Furthermore, it is assumed that a dysarthria diagnosis is established by comparing the individual's intelligibility with a threshold value established from an age-matched group with no known neurological impairment. In individuals for whom dysarthria treatment is warranted, increased intelligibility is the primary goal (Yorkston et al., 1987). Further information is then required to identify suitable targets for therapy. This can be achieved by assessing the speaker's performance over a range of speech dimensions (e.g., Darley et al., 1969a). All other factors being equal, it would be logical to devise a therapy plan for the speech dimensions that are most severely affected.

However, this assumes that all production features carry equal importance in conveying the speaker's message. In reality, some features are likely to have a greater effect than others. De Bodt et al. (2002), for example, provided evidence to suggest that *articulation* is the most relevant dimension for speech intelligibility. Yet even within the realm of articulatory imprecision, the importance of a particular error (e.g., the devoicing of word-initial consonants) is likely to depend on a variety of factors, especially *functional load* – the usefulness of the phonetic feature in conveying information within the language. These factors need to be considered when designing a tool to identify speech deficits that are worthwhile targets for therapy. For example, it would be reasonable to contend that the number of errors observed for a particular phoneme or phonetic contrast holds greater validity when obtained from an intelligibility test in which the distribution of phonemes reflects that used in everyday language.

1.3. Current approach: phonetic-contrast analysis

This thesis started from the assumption that the therapist and client have ascertained the need to carry out perceptual, articulatory analysis as a means of identifying the client's speech errors (e.g., for the purpose of selecting targets for therapy). The broad aim of the thesis was to improve understanding of the methodological issues surrounding the identification and categorisation of articulatory, segmental errors in dysarthric speakers by perceptual means. A possible approach is to use phonetic or phonemic transcription, either in a word- or sentence-reading task, or even in conversational speech (although this presents a number of challenges, including the fact that a transcript of the intended speech output may be required). However, expert transcription is a time-consuming, skilled undertaking that is unlikely to become a widespread clinical tool in dysarthria assessment. Furthermore, it does not automatically provide a means of *quantifying* errors, nor of *categorising* them according to a coherent phonetic framework. Therefore, the data would require further interpretation to identify targets for therapy.

The current thesis investigated the potential for applying the methodology proposed by Kent et al. (1989) to Belgian Dutch speakers with dysarthria. Kent et al. developed a single-word intelligibility test (for American English) that can, in principle, address all of the aforementioned limitations. The test is designed to identify errors that constitute a contrast in a single phonetic feature, such as 'stop place of articulation' or 'high vs. low vowel'. This results in a set of possible contrast errors (19 in the Kent et al. test) that are intrinsically linked to phonetic theories about consonant and vowel articulation. Kent et al.'s (1989) approach, referred to herein as *phonetic-contrast analysis*, may offer a number of theoretical and practical benefits. In particular, by limiting the outcome to a finite set of

metrics (the error rate for each phonetic-contrast category), it may be easier to (a) track the progress of an individual over time and (b) characterise and compare different dysarthria populations (defined, for example, by lesion site, aetiology, gender or severity). Therefore, phonetic-contrast analysis has the potential to be of value to clinicians and researchers alike. However, there is a lack of (cross-linguistic) evidence in support of some of the underlying assumptions of the technique.

Despite the anticipated benefits of phonetic-contrast analysis, a major limitation of the approach (or indeed of any assessment that employs single-word targets) is that it does not allow for evaluation of the *suprasegmental* properties of speech. Therefore, to obtain a complete picture of the factors affecting speaker intelligibility, further assessment would be required, e.g., the transcription of intonation by means of ToBI ("Tones and break indices"; see Beckman et al., 2005). However, it was not possible to investigate all types of speech deficit within the time frame of the current project. Therefore, this study was limited to phonemic-error analysis only. Articulatory errors are observed in all types of dysarthria (Darley et al., 1969a) and are thought to be more strongly associated with intelligibility than suprasegmental errors (de Bodt et al., 2002; Whitehill et al., 2004).

1.4. Aim and thesis outline

As stated, the broad aim of this thesis was to improve understanding of some of the methodological factors that affect the identification of articulatory errors by perceptual means in Belgian Dutch speakers with acquired dysarthria. Following a review of the literature (see Chapter 2), it was decided that the main goal would be to explore the potential for identifying errors by means of phonetic-contrast analysis (Kent et al., 1989). However, although the thesis focuses on this particular technique, the findings yield insights of a much broader nature, not only with regard to articulatory analysis in general, but also with regard to the production and perception of dysarthric speech.

The remainder of this thesis is organised as follows. Chapter 2 reviews the relevant literature, focusing mainly on the methodological aspects of identifying articulatory errors by perceptual means. Chapter 3 describes the methods common to all the investigations carried out in this thesis, such as details of the participants and the speech-production tasks. In particular, it describes the methodology behind the construction of a novel, Belgian Dutch single-word intelligibility test developed by the author. Chapter 4 describes the results of an orthographic-transcription study that aims to (a) identify phonemic errors in single-word reading for Belgian Dutch speakers with dysarthria and (b) categorise these errors according to a set of contrasts in a single phonetic feature.

Chapter 5 compares the phonemic and phonetic-contrast errors reported in Chapter 4 with the corresponding 'errors' (which are more likely to be misperceptions or manifestations of ongoing phonological change) observed in age-matched neurotypical speakers. Chapter 6 compares, for speakers with dysarthria, phonetic-contrast error profiles and word-accuracy scores obtained using two different listener-response modes: orthographic transcription and multiple choice. Chapter 7 describes a preliminary investigation that is designed to test the main premise of the thesis, namely that articulatory accuracy is an important predictor of real-world intelligibility. The study measures spontaneous-speech intelligibility in speakers with dysarthria and examines its degree of correlation with intelligibility metrics derived from single-word reading. Finally, Chapter 8 summarises and integrates the findings from previous chapters, discusses their implications, and suggests directions for future research.

2. Literature review and objectives

The starting point for this thesis was the assumption that articulatory errors play an important role in real-world intelligibility, and that the perceptual identification of such errors, along with their categorisation according to some type of theoretical framework, would be a worthwhile endeavour for many speakers with dysarthria.[1] The overarching aim was to improve understanding of some of the methodological factors affecting the *perceptual identification of segmental speech errors* in Belgian Dutch speakers with acquired dysarthria. Consequently, the main purpose of this review was to examine how the information provided by the perceptual analysis of articulatory errors is affected by methodological choices such as the speech stimuli, the listener's response paradigm, and the method of coding the errors. The review also covers two additional topics that are related to the subsidiary aims of the thesis. The first of these aims was to acquire preliminary information about the phonemic errors of Belgian Dutch speakers with acquired dysarthria. Accordingly, a section of the review describes the current state of knowledge on this topic. The second subsidiary aim was to contribute to the evidence base for the premise of this thesis – namely that articulatory errors play an important role in real-world intelligibility. Therefore, the review also provides overviews of the methods of (a) examining the relationship between specific speech errors and overall intelligibility and (b) measuring intelligibility in spontaneous speech.

The remainder of this chapter is divided into five sections. Section 2.1 examines perceptual methods of identifying and categorising segmental speech errors. Since this is the main topic of the thesis, it forms the bulk of the review. The remaining sections are related to the subsidiary aims stated above. Section 2.2 summarises the available data regarding articulatory errors in Dutch speakers with dysarthria. Section 2.3 explores methods of examining the relationship between specific speech errors and overall intelligibility. Section 2.4 reviews the most common approaches for measuring spontaneous-speech intelligibility (SSI). Finally, Section 2.5 integrates the information in the previous sections to arrive at the objectives and research questions of this thesis.

The search terms used to conduct this literature review included different combinations of the following words and phrases: *dysarthria*, *segmental errors*, *speech intelligibility*, *phonemic analysis*, *phonetic contrast*, *transcription*, and *speech articulation*. The databases that were searched included PubMed, Google Scholar and APA PsycInfo. Given that the

[1] There are people with dysarthria for whom this is unlikely to be the case, such as those who make very few articulatory errors, or those whose articulation is so severely distorted that a phonemic analysis would not be helpful. Such speakers are not considered in this thesis.

thesis addresses methodological questions, publications were selected mainly on the basis of their contribution to knowledge about *methodology* rather than their observations about articulatory errors in dysarthric speech (with the exception of the papers reviewed in Section 2.2). That is, the review focuses on studies that provide insights into the relationship between methodological choices and the nature and value of the information obtained. The evidence base on this topic is not extensive and the number of studies that had a direct bearing on formulating the current set of objectives is relatively small. As a result, rather than surveying a large body of literature, the review provides reasonably detailed descriptions of the studies that are most relevant, including sufficient information about the design choices so that the methodological implications for the current thesis are clear. Many of these papers are also cited throughout the thesis, as they provide points of comparison for some of the present findings.

As stated, the main topic of the review was the perceptual analysis of articulatory errors. While the thesis also aimed to contribute to the evidence base for the premise that articulatory errors play an important role in real-world intelligibility, the study that addressed this question was limited in scope and focused on two specific goals: (i) to measure the degree of correlation between single-word intelligibility and intelligibility in natural, unconstrained spontaneous speech, and (ii) to improve understanding of some of the methodological issues surrounding the quantification of SSI. The study did *not* aim to investigate confounding factors, such as discourse coherence, fluency or prosody, nor did it attempt to identify or investigate the use of compensatory strategies (e.g., rate reduction, clear speech, loud speech). Thus, the large body of literature on the factors that affect SSI was beyond the scope of the review, as it would not have informed the study objectives; however, a few studies are mentioned in passing and this branch of the literature is discussed in more detail in Chapter 7, to aid interpretation of the results.

2.1. The perceptual assessment of segmental speech errors

2.1.1. Rationale for the perceptual approach

The limitations of perceptual methods, such as their inherent subjectivity, have led to a growing body of research aimed at improving dysarthria assessment by instrumental means (see, for example, Kent et al., 1999; Liss et al., 2009; Kim et al., 2011; Murdoch, 2011; Rong et al., 2016). Instrumental analysis includes both physiological measures (e.g., measurement of the degree of tongue movement) and metrics derived from the acoustic signal (e.g., standard deviation of the mean frequency of phonation). These methods have the advantage that they are both quantitative and objective, and thus are ideally suited to

tracking changes resulting from intervention or disease progression. However, the outcome measures can be difficult to interpret and, in many cases, do not bear an obvious relationship to a perceptually relevant characteristic of speech. Although a number of recent studies have made important strides in addressing this shortcoming (e.g., Fletcher et al., 2017; Lansford & Liss, 2014; Rong et al., 2016), perceptual analysis by a human listener is likely to remain indispensable if the ultimate goal is to gain information about phenomena of communicative value (Howard & Heselwood, 2011). This is because the initial clinical assessment is likely to always be based on perceptual features, even if the results are then used to inform instrumental analysis (Duffy, 2005: p.11). Furthermore, the desired outcome of impairment-based therapy will always be an improvement in the perceptual characteristics of speech.

2.1.2. Perceptual assessment tools used in the clinic

The most common clinical methods for the perceptual assessment of dysarthria focus on classifying the dysarthria *type*. The first systematic classification of the acquired dysarthrias (see Table 2.1) was proposed by Darley, Aronson and Brown at the Mayo Clinic in 1969. These authors collected connected-speech samples from 212 subjects, representing seven different neurologically defined groups (Darley et al., 1969a). Within each group, there were at least 30 subjects representing a wide range of intelligibility levels. Speech characteristics were captured using a list of 36 individual perceptual dimensions, relating to pitch, loudness, vocal quality (both laryngeal and resonatory), respiration, articulation and prosody. Two overall dimensions ("intelligibility" and "bizarreness") were also assessed. Each of the 38 dimensions was rated on a 7-point scale and the results were used to identify a cluster of disordered speech characteristics for each neurological group (Darley et al., 1969a; 1969b).

To this day, the most widely used assessment tools, such as the Frenchay Dysarthria Assessment (FDA; Enderby & Palmer, 2008)[2] and the Dysarthria Profile (Robertson, 1982), are very much grounded in the Mayo Clinic approach. That is to say, these tools assess oromotor functioning, to shed light on the underlying neurological impairment, as well as all five aspects of speech production (respiration, phonation, articulation, resonance and prosody). While there may be considerable value to be gained from such assessments, their broad scope means that the information gained about any one aspect of speech production (e.g., articulation) is relatively limited. For example, the words and sentences used to test intelligibility in the FDA do not include all phonemes in all possible

[2] There is a Dutch version of the FDA, although according to Knuijt et al. (2017), it is not widely used.

positions, meaning that a thorough qualitative analysis of the speaker's articulatory errors is not possible. Furthermore, the outcome measure is the overall intelligibility (rated on a 5-point scale), which is determined from the number of intelligible test items; there is no formal framework for categorising and quantifying *specific* articulatory errors.

Dysarthria type	*Part of nervous system affected*	*Darley et al.'s (1969a) neurological group*
Flaccid	Lower motor neurones	Bulbar palsy
Spastic	Upper motor neurones	Pseudobulbar palsy
Mixed (flaccid/spastic)	Upper and lower motor neurones	Amyotrophic lateral sclerosis
Ataxic	Cerebellum	Cerebellar disorders
Hypokinetic	Extrapyramidal tract, substantia nigra	Parkinsonism
Hyperkinetic	Extrapyramidal tract, basal ganglia	Dystonia and choreoathetosis

Table 2.1. Dysarthria classification according to the RCSLT (2009). The third column shows the corresponding Darley et al. (1969a) neurological group. The groups "dystonia" and "choreoathetosis" have been merged for this table, as they both cause hyperkinetic dysarthria.

Recently, the Radboud Dysarthria Assessment (RDA), a Dutch dysarthria assessment that was originally made available in 2007, was improved and validated (Knuijt et al., 2017). The goals of the assessment are to diagnose the dysarthria type and to estimate dysarthria severity on a 5-point scale. Although the RDA places greater emphasis on the assessment of speech characteristics than does the FDA, there is no component for formal intelligibility testing. According to the authors, this is due to the fact that there are two validated Dutch intelligibility tests for these purposes, one at the word level (de Bodt et al., 2006) and the other at the sentence level (Martens et al., 2010). Of these, the word intelligibility test, the Nederlandstalig Spraakverstaanbaarheidsonderzoek (NSVO; de Bodt et al., 2006), is of greater relevance to the present study.

The NSVO requires the speaker to read 50 monosyllabic words (mainly CVC but also some CV and VC), approximately half of which are real while the other half are pseudo-words. A significant advantage of the assessment is that it randomly selects a word list from among twenty-five alternatives, such that the assessor is unaware of the targets. Furthermore, it provides information that is both quantitative (a measure of intelligibility) and qualitative (phonemic analysis). The task of the assessor is to identify the missing phoneme at one of the three word positions; i.e., the remaining two target phonemes are given. Phoneme identification is carried out using a forced-choice response mode; however, the choice

includes all phonemes that are phonologically possible at the given word position, meaning that, in effect, the test uses an open-response paradigm. Van Nuffelen et al. (2008) administered the test to 30 adults with pathological speech (11 dysarthric, 10 hearing impaired, 9 laryngectomees) and the target phonemes were identified by 9 experienced listeners. The inter-rater reliability per speaker for phoneme identification varied between fair and almost perfect (κ: 0.24-0.89), with the higher levels of agreement corresponding to speakers with greater intelligibility. The authors concluded that the technique is only reliable and clinically relevant in individuals with a mild to moderate impairment.[3] The intra-rater reliability for phoneme identification was reasonable in all speakers (κ: 0.60-0.79). Despite the advantages of the NSVO, it does not meet the requirements of the present thesis, largely due to the fact that each phoneme at each word position is only tested on one occasion. Therefore, there are insufficient data to enable a systematic investigation of error rates for different phonemes or phonetic features. In addition, the test does not include any consonant clusters, despite the fact that these occur frequently in the Dutch language. Third of all, the distribution of phonemes does not reflect that used in everyday language, which may limit the degree of correlation between the intelligibility metric and real-world intelligibility. Finally, the NSVO employs pseudo-words, meaning that it is only suitable for use with expert listeners.[4] This is because, without thorough training in responding to nonsense syllables, naïve listeners tend to respond with *real* words (Bosman & Smoorenburg, 1995).

Given the lack of an available clinical tool to meet the needs of this thesis, the first objective was a technical one – to develop a perceptual assessment that could be used to identify and categorise articulatory errors in Belgian Dutch speakers. There are two main methods of recording perceptual, segmental errors: (i) phonetic transcription, in which an expert listener transcribes the production characteristics of the perceived sound in a relatively unconstrained manner, and (ii) orthographic transcription / multiple-choice selection, where the listener (who may or may not be a specialist) is constrained to record responses that are real words or sentences of the target language. The following two subsections (2.1.3 and 2.1.4) provide reviews of these two approaches, with the goal of deciding which one would be more appropriate for the current thesis.

[3] This limitation is likely to apply to any perceptual assessment based on phonemic analysis.

[4] Although this is mentioned as a drawback of the NSVO, it was not grounds for ruling it out in the early stages of the project, as it had not yet been decided whether speech errors would be judged using expert or naïve listeners.

2.1.3. Phonetic transcription

Studies that provide a systematic perceptual analysis of segmental errors in dysarthric speech started to appear in the wake of the Darley classification system. In 1970, Johns and Darley compared the articulatory characteristics of three populations (all speakers of American English): apraxia, dysarthria and normal controls. The speakers with dysarthria demonstrated features of the following dysarthria types: spastic (n=6), flaccid (n = 2) and mixed (n = 2). A wide range of single-word and connected-speech stimuli was assessed. For the present purposes, the most relevant findings are those pertaining to narrow transcription of 60 CVC words, half of which were real and the other half pseudo-words. Only the initial consonant was systematically varied and subjected to perceptual analysis. The authors tested all consonants of English plus some common consonant clusters, and the speakers were required to produce each sound on 15 occasions. The following error types were recorded: addition, omission, substitution (with another phoneme of English), repetition, schwa insertion (i.e., splitting of consonant clusters), and distortion (i.e., the target phoneme was "recognisable but indistinct"). Cluster reductions and formations were regarded as substitutions and did not contribute to the number of omission / addition errors. The main findings for speakers with dysarthria were as follows. The most common type of error was distortion (65% of the errors). This was followed by schwa insertion (18%), substitution (9%), and addition (5%). Omission and repetition errors were rare. The six most vulnerable singleton phonemes from the point of view of phonemic errors (i.e., including substitutions, additions and omissions, but excluding distortions) were, in order of decreasing vulnerability, /r, v, b, m, dʒ, f/. The six phonemes with the lowest error rates (in fact, they yielded almost no errors) were /h, k, ʃ, t, p, g/. The full set of phonemic-substitution errors, pooled over the dysarthric group, was presented as a confusion matrix. The authors noted that, in general, speech sounds that are believed to have a higher level of articulatory difficulty (e.g., fricatives, affricates and clusters) yielded a greater number of substitutions and distortions. They concluded that the most striking feature of dysarthric speech was "phonemic simplification". They further noted that the distortions "always constituted simplifications" and that the hallmark feature was the "slighting of sounds". Furthermore, they reported that the types of articulatory error and their manner of production were highly consistent, especially when compared with speakers with apraxia.

Platt et al. (1980a, b) used a variety of tasks (single-word reading, diadochokinetic speech rate and reading of a passage) to characterise the speech of 50 adult males from New South Wales, Australia with cerebral palsy (32 spastic, 18 athetoid). The full set of single-word stimuli (mainly CVC with some VC and CV) was assessed using orthographic

transcription by naïve listeners (Platt et al., 1980b). In addition, a subset of these words was selected for expert phonetic transcription in which three classes of error were permitted: omission, substitution (with another phoneme of English), and distortion. In the case of a distortion, the following transcription markers could be employed: devoicing, nasalisation, dentalisation, aspiration, lateralisation and rounding. The inter-judge reliability of the phonetic transcriptions was assessed by obtaining transcriptions from a second phonetician for 11 of the subjects, and the mean percentage agreement was 84%. Having observed that the majority (65%) of errors in the Johns and Darley (1970) study had been classed as distortion errors, and hence were not subject to further categorisation, Platt et al. (1980b) chose to adopt a different approach. They reallocated a distortion error either to the "correct" category or the "substitution" category, depending on the level of the articulatory distortion(s). For example, a token of /v/ that was identified as a distortion, but produced with no voicing and with "usual frication" was reassigned as /f/. As a result of this process (although other factors may have played a role), Platt et al. (1980b) reported twice as many substitution errors as Johns and Darley (1970). In addition to presenting the substitution errors as confusion matrices, the authors produced a variety of summary measures from these matrices, such as the consonant manner that exhibited the most errors (fricatives) and the ratio of the frequency of within-manner substitutions to between-manner substitutions (6.5:1 and 11:1 for word-initial and word-final consonants, respectively). With regard to vowels, they noted that errors seemed to distribute in cells of the confusion matrix that were adjacent to the intended cell (i.e., vowel errors were generally confined to small shifts across the vowel space). However, the authors advised the reader to interpret the vowel errors with caution due to the fact that the phonetic transcription of vowels requires the assessor to be highly familiar with the prevailing vowel usages in the dialect in question. The final part of Platt et al.'s (1980b) study compared error profiles for speakers with different intelligibility levels (the percentage of words correctly recognised by the naive listeners). The comparison was based on two features: (1) the distribution of errors across the three phonemes (initial consonant, final consonant and vowel) and (2) the distribution of within- and between-manner consonant substitutions, depicted using a confusion matrix of "intended manner" versus "produced manner". The error profiles were similar in all three intelligibility groups, suggesting that, for this sample of speakers, the difference between severity levels was one of degree rather than quality.

Having examined two of the seminal phonetic-transcription studies in some detail, the reader should have a reasonable understanding of the nature of the information that can be gained from such an approach, as well as some of the benefits and limitations. These

points will be summarised at the end of this subsection. First, it is worth mentioning a number of other transcription studies that either (a) adopted a different approach from the previous two studies or (b) are particularly relevant to the current thesis.

Logemann and Fisher (1981) analysed the speech of 200 individuals at varying stages of Parkinson's disease (PD), 90 of whom were found to exhibit misarticulations. In contrast to the aforementioned expert-transcription studies, which used single-word targets, the test stimuli consisted of the first 11 sentences of the Fisher-Logemann Test of Articulation Competence, which tests all consonant phonemes of English in all word positions. Narrow transcription of the consonant phonemes was performed by two trained phoneticians, and inter-judge agreement regarding the identity of phonetic substitutions was high (0.93). The authors reported a high degree of consistency within and between speakers in terms of the nature of the errors. This allowed them to summarise the errors by tabulating (a) the number of speakers who misarticulated specific phonemes and (b) the most common types of misarticulation for specific phonemes. Most of the errors reflected articulatory undershoot or inadequate narrowing of the vocal tract (e.g., /k/ → [x]). The fact that articulatory undershoot is characteristic of dysarthria due to PD has been corroborated in a number of studies (Read et al., 2018).

Relatively few studies of dysarthric speech have attempted the narrow transcription of vowels, a notoriously difficult task that can result in poor inter-transcriber reliability (Howard & Heselwood, 2013). An exception is Odell et al. (1991) who studied 12 subjects equally divided among three diagnostic categories (apraxia of speech, conduction aphasia and ataxic dysarthria). The speech task consisted of single-word imitation using 30 words of increasing length (e.g., *please-pleasing-pleasingly*) taken from the Apraxia Battery for Adults. This source was chosen because it is known to elicit frequent speech errors in neurogenic populations. The word list tested 10 of the 14 monophthong vowels of American English. The phonetic transcriptions, which were produced by two experienced transcribers not involved in data collection, were based on the IPA symbols for the vowels of American English, plus the diacritics described in Shriberg and Kent (1982). The two transcribers reached a consensus transcription for each segment, following a well-known set of guidelines (Shriberg et al., 1984). The list of segmental errors included omission, addition and inaccurate productions, where the latter were categorised as substitutions, distortions or distorted substitutions. Distortions involved errors in placement, timing, or the sound source (e.g., breathy or murmured productions). Reliability of the vowel transcriptions was assessed in terms of item-by-item agreement at two different time points. For the dysarthric group, there was disagreement concerning whether or not an

error had occurred for 15% of the corpus. 81% of vowel errors in the dysarthric group were classed as distortions, with the next most common categories being substitutions and distorted substitutions (9% each). Omissions and additions were rare. The predominant type of substitution involved replacing a monophthong with a diphthong.

Whitehill and Ciocca's (2000a) study is of interest because it investigates a language other than English. The language in question (Cantonese) had received very little attention in the literature, so the authors were careful to include several exemplars (at least 3) of each vowel and consonant phoneme, with the consonant phonemes appearing at different word positions. In addition, all tones of Cantonese were tested. The test stimuli consisted of 100 monosyllabic words. The 22 speakers (12 M, 10 F), all with a diagnosis of cerebral palsy, were of varying intelligibility levels. Narrow transcription was carried out by two expert listeners and the segments were coded as follows: correct, omission, addition, substitution (not necessarily to a phoneme of Cantonese), distortion (e.g., nasalisation, dentalisation, lateralisation), or distorted substitution. Substitution errors were divided into manner only, place only, aspiration only (Cantonese does not have a voicing contrast), or a combination of these categories. Place and manner errors were further described according to the specific phonetic process (e.g., 'backing' or 'stopping'). Point-to-point reliability was calculated using the total number of segments and tones in the word list (335). Inter-judge reliability, calculated for all data across all speakers, was 92%. Discrepancies were subsequently resolved through discussion. The authors reported a wide variety of summary measures, including the relative frequencies of vowel versus consonant errors, error rates as a function of consonant place of articulation, and the relative frequencies of place, aspiration and manner errors. It is interesting to note that since substitutions were not confined to other phonemes of Cantonese, substitution was reported to be the most common error type (e.g., 70.3% of all initial-consonant errors), in contrast to the aforementioned studies (e.g., Johns & Darley, 1970; Platt et al., 1980b; Odell et al., 1991).

To summarise, the characterisation of segmental speech errors by means of phonetic transcription offers a number of advantages, but is also subject to limitations. The main advantages are applicability to a wide variety of speech stimuli (including multisyllabic words, pseudo-words and connected speech) and the ability to capture speech deficits at a range of levels, from subtle distortions to phonemic substitutions within and beyond the target language. Even when an individual is highly unintelligible, with a severely restricted production inventory, it is possible to produce a phonetic transcription of their speech, with or without knowledge of the intended utterance. When transcriptions of

dysarthric speech are carried out by experienced phoneticians, the rate of agreement between assessors appears to be high (≳ 80%), with any differences being resolved by discussion. The main disadvantages of phonetic transcription are as follows. Firstly, it is a technique that requires considerable training and experience, as well as familiarity with the normal allophonic and dialectical variations of the sample population. Secondly, the wide range of definitions used in these studies (for example: Johns & Darley (1970), Platt et al. (1980b) and Whitehill & Ciocca (2000a) all employed a different definition of a substitution error) may cause difficulties in comparing the findings; see Miller (1995) for a more detailed discussion of this issue. A related point is that there is considerable variation in the way in which authors present their findings, ranging from confusion matrices that only show phonemic substitutions within the target language (e.g., Johns & Darley, 1970) to detailed descriptions of the production characteristics of specific phonemes (e.g., Logemann & Fisher, 1981). This variation may, at least in part, reflect differences in the homogeneity of the sample population; i.e., it is only feasible to extract summary information regarding subtle distortions if the same distortions tend to arise in different speakers. This leads to the third limitation of phonetic-transcription studies: despite the fact that the technique is capable of recording subtle distortions, in many studies, the distortion errors are not reported in any detail, or if they are, they are not subjected to further categorisation. As mentioned, this could be due to the fact that there is too much variety in the nature of the distortions among speakers. It could also be due to the lack of a systematic categorisation method for distortions – as opposed to phonemic substitutions, which can be described in terms of contrasts in voice, place and manner (for consonants) or height, backness and duration (for vowels). Nevertheless, some authors have devised methods for summarising and categorising distortion errors. For example, Kim et al. (2010), in their study of speakers with cerebral palsy, coded distortions according to their phonetic nature (e.g., a lateralized /ʃ/ was categorised as a manner error to denote that the production had an unexpected liquid quality). Similarly, Haley et al. (2019) used phonetic transcription to identify errors in speakers with AOS, but then categorised the distortions in terms of articulatory deficits such as voicing ambiguity and tongue centralisation. Therefore, it is likely that the main motivation for leaving distortion errors unspecified is that it can be an extremely time-consuming endeavour, which may not be considered worthwhile if the phoneme is still recognisable as the target. This argument is even more compelling in a clinical (as opposed to a research) setting, where the primary goal is to identify functional targets for therapy rather than to improve understanding of dysarthric speech.

2.1.4. Orthographic transcription

Since phonetic transcription is a laborious process that requires considerable skill and experience, an attractive alternative is to confine listener responses to real words. It is important to understand the effect that this has on the perceived errors. In comparison with phonetic transcription, one of the main differences, of course, is that the range of errors is significantly narrowed, as it is only possible to identify phonemic substitutions within the target language. Further reduction arises due to the fact that the response is constrained to a real word such that, depending on how the other segments of the word are perceived, a given segment can only correspond to a particular set of phonemes. The difference between orthographic transcription and broad transcription was investigated by Haley et al. (2001). Their main aim was to characterise segmental speech errors in aphasia and apraxia of speech, but a section of the paper was devoted to examining the methodological issue of how broad transcription by expert listeners compares with orthographic transcription of the same speech data. They hypothesised that there would be some correlation between the two error profiles, but it would not be perfect. In orthographic transcription, listeners may hear a segmental error that does not prevent them from identifying the intended word (e.g., if the target *sheep* were perceived as /ʃib/, the listener might still transcribe the word correctly because they can only record real words). The opposite situation could also arise, with the lay listener transcribing errors that were *not* perceived in order to satisfy lexical constraints. The speech stimuli consisted of the 70 monosyllabic words from the intelligibility protocol developed by Kent et al. (1989). Broadly speaking, there was fairly close agreement between the *types* of error generated by the two approaches (categorised using Kent et al.'s list of 19 phonetic contrasts), both for the population as a whole and for individual speakers. However, the *frequency* of errors was substantially higher (a factor of almost two) for data transcribed phonemically by expert listeners. This suggests that, of the two scenarios mentioned above, the dominant process was the one in which lay listeners heard segmental errors that did not prevent them from identifying the target word.

The methodological choices in an orthographic-transcription study are relatively limited, especially if the goal is to identify articulatory errors, as in the present thesis. The most logical stimuli are real, monosyllabic words. An assessment based on the identification of monosyllabic words primarily reflects the perception of segmental, articulatory features, while the contribution of suprasegmental features to perception is likely to be negligible. Furthermore, single-word stimuli eliminate contextual cues and increase the possibility that confusion errors will arise even in speakers with mild dysarthria (who might be fully intelligible in connected speech). The available choices regarding the response protocol

are also fairly limited. The assessor needs to be unaware of the intended target (to reduce positive bias) and the method of response may consist of either transcribing the word that was perceived or selecting from a list of words that includes the target and a number of distractors. Once the responses have been recorded, they need to be analysed in such a way as to identify targets for therapy. That is, armed with the knowledge of all the substitutions made by the speaker, the assessor needs to identify which substitutions occur most often, and, if there are significant error rates for multiple substitutions, to attempt to reduce the variable space by detecting commonalities between them (i.e., substitution errors that are all indicative of the same underlying deficit, such as laryngeal or velopharyngeal dysfunction). To the best of the author's knowledge, there is only one framework in the literature (Kent et al., 1989) that endeavours to meet these needs for people with dysarthria and that has been investigated in a reasonable number of research studies. Kent et al.'s (1989) word intelligibility test, as well as yielding an overall measure of intelligibility (word accuracy), provides a means of characterising an individual's segmental errors in terms of error rates (from 0 to 1) for a set of 19 phonetic-contrast categories (e.g., 'high vs. low vowel', 'stop vs. fricative'). Thus the outcome measures are intrinsically linked to phonetic theories about vowel and consonant articulation.

2.1.5. Current approach: Phonetic-contrast analysis

The previous two subsections described, respectively, the nature of the information about segmental, articulatory errors that may be attained via phonetic and orthographic transcription. This subsection begins by comparing these two approaches in order to decide on the most suitable technique for implementation in the current thesis.

The main advantages of phonetic transcription are that it allows for the greatest amount of detail and accuracy in capturing speech characteristics, and it can be applied to any type of speech signal. However, narrow transcription requires a substantial time-commitment, meaning that the researcher may have to make compromises with regard to the population sample size and/or the range of allowed diacritics. Furthermore, since transcription requires substantial training and experience, it is often the case that the only people available to perform the task are those responsible for designing the study and collecting the data. In such cases, the transcribers may be familiar with the target words and perhaps also the neurological deficits of the speakers, which could bias their responses. Of greater concern in the present study was the lack of transcription experience on the part of the author as well as a lack of familiarity with the normal allophonic variations in the target language and accent (Belgian Dutch speakers from the Antwerp region). In addition, since this is a research study, the speech errors of each

subject would need to be transcribed by at least two listeners, in order to determine the level of inter-judge agreement, and the author did not have access to sufficient numbers of experienced phoneticians who could take on this task. The possibility of broad transcription was considered, as this has the advantage of capturing a greater number of errors than orthographic transcription (according to Haley et al., 2001), while requiring less skill on the part of the assessor than narrow transcription (meaning that a wider pool of assessors would be available). However, even with this approach, it could have still proven difficult to recruit sufficient numbers of skilled assessors. For example, Haley et al. (2001) enlisted two second-year SLT graduate students to perform phonetic transcription in their native language (American English). In addition to the standard training received during their courses, the listeners completed some narrow transcription exercises prior to data analysis. Nevertheless, when an experienced phonetician later re-transcribed the data, there was poor agreement with the original transcription, including 42% of the tokens disagreeing at the level of *broad* transcription.

Based on the above argumentation, it was decided that the current study would identify articulatory errors using an approach that would be suitable for naïve listeners: either orthographic transcription or multiple choice (see Section 2.1.6 for a comparative review of these approaches; based on this review, it was decided that *both* methods would be investigated in the current thesis). The advantage of using an approach in which the listener's response is confined to a real word is that all perceived distortions correspond to a reduction in intelligibility and hence have functional relevance. In contrast, phonetic transcription requires further interpretation to determine which of the perceived errors are likely to have consequences for intelligibility.

Having decided to use naïve listeners, the remaining methodological choices, as argued in Section 2.1.4, are somewhat limited. The most logical choice of stimulus is a set of real, monosyllabic, single words. Regarding outcome measures, the only systematic framework for quantifying and categorising phonemic errors that has been given serious consideration in the literature is Kent et al.'s (1989) method of *phonetic-contrast analysis*. The remainder of this subsection is devoted to critiquing this technique which, despite having been implemented in a reasonable number of research studies, has not yet been subject to rigorous investigation to determine the validity of its underlying assumptions.

Kent et al.'s (1989) test consists of 70 monosyllabic, real words, chosen to be minimally contrastive with a large number of other words. The listener records their response using a multiple-choice protocol: the target word plus three foils (e.g., <u>*witch*</u> - *wit rich wish*), where each foil represents a contrast in a single phonetic feature of one of the three

phonemes. The outcome of the assessment is a distribution of error rates across 19 phonetic-contrast categories (see Table 2.2), which can be displayed as a bar chart.

Label	*Phonetic contrast*	*Word pair example(s)*	*# potential errors*
1	Front-back vowels	*feed – food*	11
2	High-low vowels	*feet – fat*	12
3	Long-short vowels *	*beat – bit*	11
4	Voiced-voiceless consonants (word-initial)	*pat – bat*	9
5	Voiced-voiceless consonants (word-final)	*bad – bat*	11
6	Alveolar-palatal fricative	*see – she*	8
7	Other fricative places of articulation	*sigh – thigh*	15
8	Stop and nasal place of articulation	*cake – take* *meat – neat*	9
9	Fricative-affricate	*ship – chip*	9
10	Stop-fricative	*sip – tip*	20
11	Stop-affricate	*chop – top*	6
12	Stop-nasal	*beat – meat*	10
13	Glottal-null (syllable-initial [h] vs. null)	*hand - and* *air – hair*	11
14	Initial consonant-null	*air – fair* *sink – ink*	14
15	Final consonant-null	*rake – ray* *blow – bloat*	9
16	Initial cluster-singleton	*steak – take*	12
17	Final cluster-singleton	*sticks – stick* *rock – rocks*	12
18	[r]-[l]	*rock – lock*	10
19	[r]-[w]	*read – weed*	8

* This is a convenient label, as the vowels in *beat* and *bit* differ in both duration and quality.

Table 2.2. Kent et al.'s (1989) list of 19 phonetic-contrast categories. The number of potential errors (final column) reflects the number of occasions on which the confusion is possible in their multiple-choice test. For some items, two of the distractors test the same phonetic contrast (e.g., *steak* – *snake sake take* yields two opportunities for initial-cluster reduction). In such cases, the authors counted this as two potential errors.

The viewpoint taken in this thesis is that the phonetic-contrast categories are named after the substitution identified by the *listener* (e.g., 'initial singleton vs. cluster'). It appears that Kent et al. (1989) assumed that the perceived error was also an accurate descriptor of the nature of the misarticulation. For example, in a follow-up paper (Kent et al., 1990), the authors present a table (Table 2) showing the articulatory correlates of their categories. They describe initial singleton vs. cluster errors as "production of a single consonant vs. a sequence of consonants in syllable-initial position." Yet, it cannot simply be assumed that the *perceived* error corresponds to the *produced* error, as the perception of a cluster may arise because the speaker has distorted the target singleton, and not because they inserted a phoneme (i.e., a true "intrusion" error). Therefore, as mentioned, the current thesis uses the labelling system of Kent et al. (1989) as a means of providing an unambiguous description of the perceived error, without intending to make any assumptions about the underlying cause (which, in some cases, may be wholly perceptual, e.g., confusions between phonemes of low perceptual distinctiveness). More generally, throughout this thesis, unless stated otherwise, the term "error" is used to describe the token perceived or transcribed *by the listener*; it does not imply that the speaker produced the substitution in question or indeed a misarticulation of any kind.

The error rate for a given category (a ratio from 0 to 1) is calculated as the number of detected errors divided by the number of potential errors based on the distractors. The concept of an error rate of this kind, where the number of potential errors is known, is referred to herein as "vulnerability". Kent et al. (1989) chose their word list and distractors so as examine 19 phonetic-contrast categories that they claimed to be vulnerable in dysarthria. Evidence for this was obtained from a review of the perceptual and acoustic studies of dysarthric speech available at the time.[5] A further stated criterion for choosing the contrasts was that each one should be capable of being captured by one or more acoustic metric(s). The rationale for this decision was that, having identified an individual's vulnerable phonetic contrasts, the researcher or therapist can subject these contrasts to further analysis based on acoustic measures.

Having introduced their intelligibility test, Kent et al. (1989) illustrated its application to a group of 13 male speakers with amyotrophic lateral sclerosis (ALS) who had different levels of dysarthria severity. The authors showed that a discernible pattern of commonly affected phonetic contrasts did emerge and that it differed slightly between groups defined by overall intelligibility (word accuracy). They further noted that the vulnerable

[5] The evidence base for Kent et al.'s (1989) list of contrast categories is also discussed in Weismer and Martin (1992).

features were mainly associated with velopharyngeal function and laryngeal configuration, which they argued to be consistent with perceptual descriptions (specifically, hypernasality and harshness) of the speech of individuals with ALS.

Kent et al.'s (1989) test was intended to be used as a clinical tool and the authors provided some guidance as to how this might be achieved. In particular, multiple, parallel word lists (ideally randomly generated) would be required to reduce listener (and, for serial assessments, speaker) familiarity. However, the authors themselves did not produce such lists. This limitation, combined with the fact that the test has not been validated or standardised, may explain why it is not used in the clinic. However, it has been implemented in a number of research studies, the purpose of which was to identify the contrast-error profiles of subject groups defined by aetiology (see, for example, Blaney & Hewlett, 2007; Bunton & Weismer, 2001; Haley et al., 2000; Kent et al., 1990; Kent et al., 1992; Bunton et al., 2007).

Before describing the advantages and shortcomings of Kent et al.'s (1989) approach, it is worthwhile defining the relevant terminology used in this thesis. The term *phonetic-contrast analysis* refers to the principle of encoding phonemic errors according to a set of error rates for a finite number of predefined categories, each of which represents a contrast in a single phonetic feature. Furthermore, it is assumed that the contrast categories in question are defined in a similar way to that suggested by Kent et al. (1989) for American English (see Table 2.2), i.e., according to the descriptive features of vowels and consonants commonly used in articulatory phonetics.[6] This definition of *phonetic-contrast analysis* only encompasses the method of error analysis; it makes no assumptions about how the data were collected. For example, as mentioned in Section 2.1.4, Haley et al. (2001) used phonetic-contrast analysis to categorise segmental errors obtained using broad transcription. However, a phonetic-contrast assessment based on a *closed* response paradigm would present a significant advantage, as it would accelerate the process of transcription (i.e., the speed with which the listener can record their responses) and, even more so, the process of error coding. Therefore, the use of a multiple-choice paradigm can be considered a desirable feature of phonetic-contrast analysis in the sense that the method might be considerably less useful, at least in a clinical setting, if the assessor were required to identify and categorise the errors themselves. Nevertheless, to avoid confusion, in this thesis the term *phonetic-contrast analysis* refers solely to the method of coding, while a reference to the approach of *Kent et al. (1989)* assumes, in addition, that

[6] Naturally, the exact set of categories will be language-dependent; for example, Category 5 in Table 2.2 is not applicable to Dutch, as all word-final consonants are devoiced.

the phonetic-contrast errors are identified and categorised automatically, through the use of a multiple-choice response paradigm.

The remainder of this subsection critiques Kent et al.'s (1989) approach at two different levels. Firstly, the review examines the implications of some of the specific methodological choices made by Kent et al. in the design of their intelligibility test. This discussion will be used to inform the process of developing the novel, Belgian Dutch single-word intelligibility test employed in the present thesis. The second part of the review discusses the advantages and limitations of the general principle of characterising speech errors using phonetic-contrast analysis.

Close examination of the word list and foils proposed by Kent et al. (1989) reveals a number of shortcomings of their test design, which, to the best of the author's knowledge, have not been mentioned by previous researchers. Some of these limitations seem to have important implications when implementing the test in a clinical or a research setting.

- The list of target words is limited in its phonemic inventory. Examples of consonant phonemes that appear in the target words, in a contrastive position, on fewer than three occasions include /g, m, n, v, w, z, θ, ð, ʒ, dʒ/; in fact, the voiced fricatives do not appear at all. Given the many constraints on the word list, it may not have been possible to include multiple examples of all phonemes of English. Alternatively, Kent and colleagues may have limited the use of certain phonemes, on the grounds that they are thought to be less vulnerable in dysarthria.[7] Nevertheless, it is important to be aware that some contrast categories are more specific than their name implies.
- The set of contrasts and contrast categories tested by Kent et al. (1989) does not seem to be exhaustive (or even optimal) based on some of the findings in the research literature. For example, Odell et al. (1991) observed that the predominant vowel substitution in American English speakers with ataxic dysarthria involved replacing a monophthong with a diphthong – a contrast that does not appear in Kent et al.'s list. Furthermore, in Platt et al.'s (1980b) study of Australian speakers with cerebral palsy, the vowel errors typically involved confusions between phonemes that are relatively close together in F1-F2 space (for normal speakers); thus, the *feed* – *food* substitution, which is the most common minimal pair used to test the 'front-back vowel' category in Kent et al.'s assessment, was not observed on even one occasion.[8]

[7] In Kent et al.'s (1989) literature review of error profiles in speakers with dysarthria, many of the observations relate to specific phonemes.

[8] This is not to say that the *category* 'front-back vowel' should be excluded. However, it may be the case that only minimal pairs that constitute a small difference in backness will yield errors.

- For some contrast categories, the word list only allows for errors in one direction (or predominantly one direction). For example, there are many more opportunities for perceiving /r/ as /w/ (e.g., *rise* - *wise*) than vice versa. No rationale for these asymmetries was provided. It is possible that Kent et al. were influenced by previous studies. For example, Johns and Darley (1970) observed many more /r/ → /w/ substitutions than vice versa. However, some of the omissions do not appear to be justified. Consider the 'stop-fricative' contrast category. The opportunities for errors are all in one direction (*sip* – *tip*), despite the fact that the frication of stops is reported to be very common in Parkinson's disease (Logemann & Fisher, 1981). Therefore, the Kent test is unlikely to capture the full range of contrast errors in some speakers.
- For some of the consonant categories, the position of the target phoneme is explicitly mentioned (e.g., 'word-initial voiced-voiceless consonants'). For others, however, the position can only be deduced by examining the word list, which reveals that the pre-vocalic and post-vocalic positions are not tested with equal frequency. Researchers who use Kent et al.'s test may not be aware of this asymmetry; yet it can be important. For example, in their study of speech errors in aphasia and apraxia, Haley et al. (2001) reported that the consonant error profile varied depending on the position of the target phoneme (pre- or post-vocalic) and that only errors in the post-vocalic position were significantly correlated with overall intelligibility.
- There are numerous other linguistic and phonetic variables (both of the target words and the distractors) that could influence listener response, but have not been discussed by Kent et al. (1989). For example, test items that simultaneously test vowel and consonant contrasts (e.g., *pit* – *pet pat bit*) could bias the observer towards choosing the consonant contrast when both errors are produced at the same time (e.g., /bɛt/). This is because listeners tolerate greater phonetic variation in vowels than in consonants before they perceive a different phoneme (Haley et al., 2000). Similarly, items that pit a pre-vocalic contrast against a post-vocalic contrast (e.g., *bad* – *bed bat pad*) could be considered biased because initial consonants are more easily identifiable than final consonants (Redford & Diehl, 1999).

Clearly, it would not be possible to control for all of the above variables and still meet the other requirements for the design of a multiple-choice intelligibility test of the type developed by Kent et al. (1989). Nevertheless, it is important to be aware of the numerous factors that can affect the listener's response. The remainder of this review focuses on the benefits and limitations of phonetic-contrast analysis as a method of encoding segmental speech errors. The main benefits of phonetic-contrast analysis stem from its structured approach. These can be summarised as follows (see Kent et al., 1989 for a more detailed

discussion): (1) It provides a coherent, evidence-based model for quantifying and categorising dysarthric speech errors; (2) The model is rooted in a theoretical framework (articulatory phonetics); (3) The approach facilitates consistency across different research studies (or, in a clinical setting, across assessments carried out at different time-points); (4) The findings are amenable to further exploration by means of acoustic analysis; (5) All the errors involve phonemic substitutions such that the errors are inextricably linked with word intelligibility. Despite these benefits, the utility of the information provided by phonetic-contrast analysis to the clinician or researcher is subject to a number of limitations and rests on a number of untested assumptions. The remainder of this subsection addresses these issues.

Lack of evidence for the validity of the categorisation method. One of the main assumptions of phonetic-contrast analysis is that the range of phonemic substitution errors typically produced by speakers with dysarthria (with the exception of speakers of very low intelligibility – see Footnote 1) is adequately represented by a reasonable number ($\lesssim 20$) of phonetic-contrast categories. This assumption has not been widely investigated in the literature. Furthermore, the little evidence that does exist is not applicable to the present population (Belgian Dutch speakers with dysarthria). Nevertheless, it is worth describing the findings of the most relevant studies, as they may shed light on the design of the current research protocol. Miller (1995) categorised errors identified by narrow transcription for a task in which 51 words (some of them polysyllabic) were elicited by picture naming and word repetition. The sample consisted of 27 subjects, five of whom had been diagnosed with dysarthria, while the remainder had either speech dyspraxia or phonemic paraphasia. Additionally, some of the subjects without dysarthria had language dysfunction. The error categories included distortions, omissions, additions, displacements (i.e., anticipatory, perseveratory or transposition errors), and non-displacement substitutions. Some of these categories were separated into further subcategories. Of particular interest for the present study are the findings for the category consisting of non-displacement substitutions: while substitutions across a categorical boundary (e.g., *desk* → [tɛsk]) were relatively common, comprising 272 out of the 1736 segmental errors observed across all speakers, more distant substitutions (referred to as substitutions with no apparent source in the sound environment), such as *desk* → [mɛsk], comprised only 13 errors in total.

Haley et al. (2000) likewise reported that the proportion of far-from-target substitutions was relatively small in their study of American English speakers who had aphasia with and without coexisting apraxia of speech (AOS). The speakers produced 100 monosyllabic

words from Kent et al.'s (1989) test – the 70 target words plus 30 additional words taken from the multiple-choice foils. The perceived words were recorded via orthographic transcription, and Kent et al.'s list of 19 phonetic contrasts was then used to categorise the errors in all three word segments. Haley et al. (2000) found that the proportion of errors that did *not* conform to Kent's list (coded as 'other') did not exceed 15% in any aetiological group.[9] Some of these errors (e.g., 'central vs. non-central vowels' and 'monophthongs vs. diphthongs') could have been encoded by expanding the list of phonetic-contrast categories. However, the majority of uncoded errors consisted of confusions across more than one phonetic feature simultaneously (e.g., *feet* perceived as *meet*). It is worth noting that, according to Haley et al.'s (2000) coding rules, some types of multiple-contrast confusion were recorded as errors in multiple categories rather than being confined to the 'other' category. Specifically, they allowed errors in vowel duration and consonant voicing to be coded at the same time as one other phonetic-contrast confusion. The authors did not explain the rationale for this decision, but presumably, these particular multiple confusions arose too often to be omitted from the analysis. The implication is that if these rules had not been invoked, the proportion of errors that would have been confined to the 'other' category would have been greater than 15%. In addition, the authors stipulated that vowel confusions involving both frontness and height should be coded as a front-back confusion only. Again, the rationale was not provided, but it can be surmised that pure errors in vowel frontness were not common (i.e., they were often accompanied by a vowel-height error) and/or vowel-height confusions were considered 'normal' rather than an indication of a speech deficit; indeed, the 'high vs. low vowel' category yielded the highest error rate of all categories in the control group, although the error rate was still lower than that observed in speakers with aphasia and aphasia + AOS.

While Haley et al.'s (2000) findings are not directly relevant to the current study (in terms of language and type of speech deficit), they provide some useful insights. In particular, the fact that some of the uncoded errors could have been accounted for, by including additional contrast categories, suggests that it may be possible to code such errors in the present study by analysing the data using an *open* response mode in the first instance. Thus, rather than making assumptions about the relevant contrast categories, these will be chosen so as to capture the full range of common confusions perceived in the cohort. This two-stage approach was also adopted by Whitehill and Ciocca (2000a; 2000b) in developing a version of Kent's test for Cantonese speakers with cerebral palsy, where the

[9] Note, however, that since prevocalic, vocalic, and postvocalic segments were analysed separately, the proportion of *words* that would have contained more than one phonetic-contrast error (i.e., when summing over all three segments) would have exceeded 15%.

final set of contrast categories was chosen based on errors identified by narrow transcription. The second type of error that went uncoded in Haley et al. (2000), namely errors that consist of multiple phonetic contrasts on a single segment (e.g., *feet – meet*), presents a greater challenge. If such errors arise frequently, then, in the case where an open response mode is used, the process of error coding would become very laborious, especially if multiple confusions are observed on multiple segments. In the case of a closed response format (which is argued to be the preferred method of implementing the technique; see Section 2.1.5), it may not be possible to include all instances of multiple contrasts among the foils.[10] It remains to be seen whether multiple-contrast errors are a common occurrence in Belgian Dutch speakers with dysarthria. It certainly seems likely that some contrasts will be sufficiently vulnerable such that they often co-occur with another type of confusion. It was surmised above that Haley et al. (2000) found this to be the case for American English speakers with regard to vowel duration and consonant voicing. Given that confusions between voiced and unvoiced word-initial consonants are not uncommon for neurotypical Dutch speakers (Pols, 1983), it would not be surprising if voicing errors were found to occur frequently in speakers with dysarthria in conjunction with manner or place errors. It also seems highly likely that Dutch speakers will, at least for some of the targets, produce errors on multiple segments.

Lack of evidence concerning intra- and inter-rater agreement. Another aspect to be considered when proposing a new scheme for coding speech errors is whether the outcome measures have high intra- and inter-rater reliability; i.e., whether, for a given speaker, a similar distribution of phonetic-contrast errors would be yielded by (a) the same listener on different occasions and (b) different listeners. An investigation of intra-listener agreement was considered to be beyond the scope of the present thesis, as there were insufficient resources to repeat any of the stimuli presented to the listeners. Therefore, the following summary focuses mainly on inter-rater agreement. Knowledge of this variable is particularly important if the goal is to use the technique in the clinic (e.g., to select intervention targets), where it would not be practical to compute an average error profile across multiple assessors. Few studies have reported inter-rater reliability values directly, at least not for a task that is relevant to the present study, where the listener responds by identifying a real word, and the degree of inter-rater agreement is calculated for *qualitative* information (i.e., phonemic or phonetic-contrast errors) rather

[10] This would depend on the range of responses observed. For a given target, provided the *total set of words* perceived across all speakers is relatively small, it may not matter if some of these words constitute errors in multiple phonetic contrasts and/or on multiple word segments. In other words, the foils could consist of words that vary in terms of their phonetic distance from the target.

than for a quantitative measure of intelligibility.[11] Bunton and Weismer (2002) measured the reliability of categorical assessments for two of Kent et al.'s (1989) contrast categories ('initial stop voice' and 'glottal versus null'). The speakers consisted of 25 individuals with dysarthria and 10 normal controls. The authors found strong inter- and intra-rater agreement (0.79 and 0.86, respectively) for assessments carried out by two of the investigators. Lillvik et al. (1999) found low inter- and intra-judge agreement for the articulatory analysis of 9 dysarthric speakers by three qualified clinicians based on a Swedish word-intelligibility test with a multiple-choice response format. The level of agreement increased with greater intelligibility, suggesting that the selection of intervention targets based on segmental analysis of an intelligibility assessment may only be feasible in speakers with a certain degree of intelligibility. Kim et al. (2010) measured intra-listener (but not inter-listener) variability based on orthographic transcription for single words uttered by seven American-English speakers with cerebral palsy. The listeners had no more than incidental experience of people with speech disorders. In addition to transcribing the perceived word, the listeners were instructed to rate their level of certainty on a three-point scale (0 – *not sure at all*; 2 – *completely sure*). The intra-listener reliability (measured by repeating the assessment for a subset of the word list) was 58.6% for words marked '1' or '0', implying that the *inter*-listener reliability for such words would have been even poorer. However, it is worth noting that in cases of disagreement, the two transcriptions were phonetically similar (e.g., *sink* and *think* for the word *sentence*). Thus a measure of reliability based on the constituent phonemes (rather than the whole word) would have yielded more favourable findings. The study that is likely to be most relevant to the present thesis is that of van Nuffelen et al. (2008), who examined the reliability of segmental analysis conducted using their proposed Dutch intelligibility assessment (de Bodt et al., 2006). Recall that this assessment is akin to an open response mode in the sense that the listener chooses from among all possible phonemes at the given word position. Therefore, the *a priori* chance of agreement is much lower than in a typical closed-response paradigm. The subjects, who represented a wide range of severity levels, comprised three subgroups: 11 individuals with dysarthria due to various aetiologies, 9 individuals with laryngectomees, and 10 with impaired speech secondary to hearing impairment. The sample of listeners consisted of speech and language therapists with at least three years' experience of the speech pathologies. The intra-rater reliability for exact phoneme identification was good (κ: 0.603 - 0.787 across

[11] Inter-rater agreement on the nature of the transcription is often reported to be high in narrow-transcription studies (see Section 2.1.3). However, the latter is not a comparable task for a number of reasons, including the facts that: (1) it allows for the coding of distortions; (2) it tends to be conducted by experienced phoneticians; and (3) the transcribers are often aware of the target.

the cohort). The level of inter-rater agreement varied considerably, depending on the speaker: the value of κ ranged from 0.30 to 0.81 for word-initial consonants, from 0.32 to 0.89 for word-final consonants, and from 0.24 to 0.84 for medial vowels. In common with Lillvik et al. (1999), van Nuffelen et al. (2008) found that the higher the intelligibility of a speaker, the higher the level of inter-rater agreement. Therefore this positive correlation appears to be consistent across different languages, assessment tools and speech pathologies. The conclusion from the above studies is that inter-rater agreement is only likely to be acceptable for speakers with a particular level of intelligibility, especially for the open response format, which is likely to show poorer agreement than the multiple-choice method (see Vigouroux & Miller, 2007, although note that their findings relate to inter-rater variability of overall intelligibility scores, not of phoneme identification). Based on this information, it was decided that the current thesis would exclude speakers with severe dysarthria, to improve the chances of reasonable inter-rater agreement. Nevertheless, it might still be the case that the inter-rater variability is too large for the technique to be clinically useful, and one of the aims of the thesis was to examine this question.

Lack of normative data. A further implicit assumption of phonetic-contrast analysis is that the observed errors are always indicative of disordered speech production secondary to dysarthria. Yet it could be the case that 'errors' would also be observed in age-matched neurotypical speakers due to, for example, the perceptual similarity of particular pairs of phonemes or a reduction in intelligibility associated with normal aging (e.g., due to muscle weakening or slower movement of the articulators). Kent et al. (1989) did not provide any normative data for their assessment, but a number of other authors have measured accuracy for single-word reading tasks in a control population. In Bunton and Weismer (2001), the main goal was to study the acoustic correlates of vowel-height errors (as perceived by human listeners) in a variety of clinical populations. However, the authors also administered Kent et al.'s test to their participants, in its original format (i.e., with a four-alternative forced-choice response paradigm). The neurotypical control group consisted of 5 males and 5 females (mean age 71.4), all with an Upper Midwest accent of American English. The mean word accuracy for the control group, assessed by 10 undergraduate students enrolled in a communication disorders programme, was 96.4% (range 94.4 - 98.4). The most common errors, which differed with sex, are shown in Table 2.3. Error rates for specific contrast categories were not reported; however, given the word accuracy values stated above, these would have been very low.

Female	*Male*
High vs. low vowels	High vs. low vowels
Voiced /voiceless initial consonant	Initial /h/ vs. null
Stop vs. nasal	Long vs. short vowels
Other fricative place of articulation	Stop vs. nasal
Final cluster vs. singleton	

Table 2.3. The top contrast-error categories (see Table 2.2 for their definitions) reported by Bunton and Weismer (2001) for neurotypical speakers of American English.

Johns and Darley (1970) also reported high single-word intelligibility for their control group of American English speakers (n = 10). Using broad transcription, they obtained a word accuracy value of 99% for a list of thirty real, CVC words. However, it is worth noting that only the initial consonant was systematically varied in this word list, meaning that vowel and final-consonant errors could not be detected; as can be seen from Table 2.3, vowel errors, in particular, seem to be among the more common type of error observed in neurotypical speakers (see also Haley et al., 2000). Furthermore, Johns and Darley's control group was not ideal with respect to demographic features (8 men and 2 women with a mean age of 36 and a range from 19 to 58 years). Odell et al. (1991) detected very few misarticulations in their four control speakers of General American English (all male, age range 57 to 67 years). The authors only reported vowel errors, which they identified using narrow transcription. The target stimuli were words of increasing length. In total, there were just two vowel errors across the four subjects, both of which were distortions (i.e., no substitutions). However, Odell et al.'s (1991) findings may be of limited relevance to the present study, as the target words were less confusable and were (presumably) known to the transcribers.

Other studies have reported higher error rates, at least in some individuals. In Vigouroux and Miller (2007), the authors devised their own list of target words, adapted to the sounds and contrasts of speakers from North East England (their sample population). The control group consisted of 24 speakers (12 M, 12 F) with a mean age of 68 years (range 43 to 77). The listeners, who comprised 40 university students and 21 older people recruited from a local voluntary organisation, had no previous experience of communicating with someone with acquired dysarthria. The word accuracies for the forced-choice response mode[12] were substantially lower than in Bunton and Weismer (2001), with a mean of 91.7% and a range of 79.3% - 97.7%. This difference may have been largely due to the

[12] The authors also used an open response mode, which produced even lower word accuracies, with a mean of 81.9% and a range of 56.7% to 95.5%.

greater number of foils used by Vigouroux and Miller (11 instead of 3), but it is likely that other factors also played a role, such as the specific word list, the accent of English, and the skills and experience of the listeners. Haley et al. (2000) used orthographic transcription to identify words uttered by 10 neurotypical speakers of American English (6 M, 4 F) with a mean age of 62 (range 50-72). The stimuli consisted of the 70 words from Kent et al.'s (1989) test. The listeners appeared to be non-experts (although this is not explicitly stated), ranging in age from 22 to 52 years. The word accuracy, obtained by averaging over ten listeners per speaker, ranged from 88.5% to 97.8%, with a mean of 95.2%. Despite the use of a fully open format, these intelligibility values are higher than those reported by Vigouroux and Miller (2007), again suggesting that differences in the word list, accent of English, and sample populations (both speaker and listener) may yield different results. The suggestion that accent is important is reinforced by the fact that in Haley et al. (2000), 41% of contrast confusions perceived for the control group were classified within the two categories 'vowel height' and 'vowel duration'. Given that vowels are much more variable between accents of English than consonants, it would not be surprising if the vulnerability of vowel contrasts were likewise accent-dependent.

In summary, the findings for word accuracy in English speakers with no known neurological impairment are mixed, but there is certainly enough evidence of relatively low accuracies (< 90%) in some speakers to suggest that normative data need to be acquired in tests of this type. Regarding the Dutch language, to the best of the author's knowledge, there are no normative data for an intelligibility test involving highly contrastive, monosyllabic, real words. However, the test developed by de Bodt et al. (2006) has been administered to a control group of 48 female and 33 male Belgian Dutch speakers (Xue et al., 2019). The results, reported in terms of the proportion of correct phonemes, ranged from 82% to 100%, with a mean (± 1 SD) of 94.2% (± 4.2%). In a study by van Nuffelen et al. (2009b), also using the de Bodt et al. (2006) assessment, the phoneme intelligibility of the 51 control speakers ranged from 84% to 100% with a mean of 93.3%. These findings reinforce the notion that normative data would be helpful in the present thesis – not just as a means of determining a threshold for dysarthria detection using single-word intelligibility tests, but also to establish whether specific contrast categories should be excluded from a dysarthria assessment because they are equally vulnerable in neurotypical speakers. Indeed, this is a distinct possibility given the phonological processes that are currently taking place within Dutch-speaking communities. For example, word-initial fricative devoicing has been reported in a variety of accents of Dutch, including Belgian Dutch (Verhoeven & Hageman, 2007), and the

Antwerp accent has very similar average formant values for [i] and [ɪ] as well as for [a] and [ɑ] (Verhoeven, 2005).

Limited understanding of the relationship with real-world intelligibility. A further implicit assumption of phonetic-contrast analysis is that a representation of impaired intelligibility based on such errors is a worthwhile endeavour, i.e., that the number and types of error identified by phonetic-contrast analysis are predictive of real-world intelligibility. Yet, as discussed further below (see Section 2.3), the ecological validity of intelligibility measures derived from single-word reading may be called into question. This is especially true for a test that is designed to measure error rates across a variety of contrast categories with approximately equal reliability, because contrast categories that do not have a high functional load in the language (e.g., the alveolar-palate fricative) make a disproportionate contribution to the overall intelligibility measure.[13] In addition to the fact that the overall intelligibility metric derived from phonetic-contrast analysis may not be indicative of intelligibility in spontaneous speech, it cannot simply be assumed that a particular phonetic deviation that is prominent in single-word reading has a substantial effect on real-world intelligibility. Firstly, the deviation in question may be consistent, predictable, and/or of low importance for linguistic judgements (Haley et al., 2000). Secondly, compensatory strategies used during connected speech (e.g., a reduced speech rate or increased loudness / effort) may alter the frequency and distribution of segmental errors relative to that produced in single-word reading.

Conflation of error categories. The final limitation considered in this review is associated with the fact that phonetic-contrast categories of the type proposed by Kent et al. (1989) might conflate errors that are due to different underlying articulatory deficits, thereby reducing the potential value of the assessment. In particular, Kent et al.'s outcome measures do not provide information about the error *direction*. For example, the category 'nasal vs. plosive' would conflate instances of nasals produced as plosives (hyponasality) and plosives produced as nasals (hypernasality), despite the fact that there could be considerable value in knowing which of these deficits is present. In some senses, this is not a fundamental critique, as it would be relatively straightforward to calculate and visualise (e.g., using a stacked bar chart) individual error rates for the two directions.

[13] As can be seen from Table 2.2 (final column), the number of potential errors in Kent et al.'s (1989) test did in fact vary across the contrast categories. In another publication from the same research team (Weismer & Martin, 1992: p.110), it was explained that this distribution was "loosely" chosen to reflect the approximate frequency with which the phonetic contrasts were thought to occur in English. However, inspection of Table 2.2 reveals that the distribution is still far from realistic, with many contrasts over-represented compared to their relative frequency in the English language.

However, if it is shown to be important to break down each category into error rates for the two possible directions, then the technique becomes less feasible because the greater the number of categories for which an error rate needs to be calculated, the larger the word list and the more time-consuming the assessment. This point is illustrated by Blaney and Hewlett (2007), who took account of error directions when they applied Kent et al.'s (1989) test to 11 male subjects from Northern Ireland with Friedreich's ataxia. The first noteworthy point regarding their methodology is that each subject was required to read 96 words, which is 26 more than the number appearing in Kent et al.'s original list. The authors did not provide their word list; however, the examples mentioned are all from Kent et al. (1989), suggesting that the additional words were generated by using some of the Kent foils as targets. It appears that Blaney and Hewlett included the extra words as a means of creating a more even distribution of error *directions*. This can be gleaned from the fact that, for their 'final-plosive voicing' category, they stated that there were 7 voiceless and 6 voiced targets. In comparison, Kent et al. (1989) included 8 voiceless, but only 3 voiced targets. The six highest error rates across Blaney and Hewlett's cohort of speakers occurred for the categories listed in Table 2.4. Directional errors were calculated as in the following example (see Column 1): on 24% of the occasions where a word-final voiced plosive could be misperceived as voiceless, this actually occurred, whereas only 4% of word-final voiceless plosives were misperceived as voiced.

Error ranking	**1**		**2**		**3**		**4**		**5**		**6**	
Error category	Final plosive voicing		Initial /h/ vs. null		Stop and nasal place*		/r/ vs. /w/		Final consonant vs. null		Initial plosive voicing	
Target phoneme	v+	v-	glottal	null	oral	nasal	/r/	/w/	Final cons.	null	v+	v-
% error	24	4	11	9	5	8	6	3	6	4	2	7

* For this category, errors were not subdivided by *direction* (fronting vs. backing) but by consonant *manner*. Thus place errors are more frequent in nasal stops than in oral stops.

Table 2.4. The six most frequently-observed error categories in Blaney and Hewlett (2007). For each category, the table also shows the percentage of occasions on which errors in each *direction* occurred (final row). The abbreviations v+ and v- denote voiced and voiceless, respectively.

The results indicate that some of the categories (especially those involving voicing) exhibited a predominant error direction, suggesting that an assessment that does not distinguish between errors in different directions would be less clinically useful. A further point of interest is that Blaney and Hewlett chose to break down the 'stop and nasal place' category (Column 3) into 'stop place' and 'nasal place' errors rather than 'fronting' and

'backing' errors (for both stops and nasals). The authors did not provide their reasoning, but it could stem from the fact that, even in speakers with no known deficit, the place characteristics of nasals are thought to be difficult to perceive (Narayan, 2008; Black, 1969). Therefore, oral place errors might have been considered more indicative of a true production error (as opposed to a misperception) than nasal place errors. Irrespective of their reasoning, the refinements introduced by Blaney and Hewlett (2007) raise the question as to whether, in addition to error *direction*, there are other distinctions that have been neglected by the Kent categorisation and that might have diagnostic value. For example, it could be the case that the frequency of occurrence of a given type of error (e.g., the frication of a plosive) is dependent on the *specific phonemes* in question, and that this differential behaviour has diagnostic value. Another example is that some of the consonant contrasts may show a different distribution of error rates depending on whether the contrast applies to the word-initial versus the word-final segment. Therefore, the data from the orthographic-transcription study will need to be carefully inspected to ensure that the process of reducing the errors to phonetic-contrast categories does not mask potentially useful distinctions. If the number of categories becomes too large, then this calls into question the feasibility of the approach, as it would require an extensive word list to measure the error rates for a large number of contrast categories with reasonable reliability. In other words, the advantage of the technique lies in its ability to reduce the speaker's errors to a relatively small number of categories, each of which provides unique information. If such a reduction is not justified, then the technique has little benefit over an assessment of a speaker's phonemic inventory, such as the NSVO (de Bodt et al., 2006). Essentially, the critique here is the same as that raised at the beginning of this discussion: it has yet to be shown that the range of phonemic-substitution errors perceived in speakers with dysarthria can be adequately represented by a reasonably small number of phonetic-contrast categories.

In summary, in order to establish whether phonetic-contrast analysis is a suitable technique for identifying targets for therapy in an individual with dysarthria, the following questions need to be answered. These questions have not been definitively answered for speakers of English, while there are almost no data for speakers of Belgian Dutch.

- *Is the range of phonemic-substitution errors typically observed in speakers with mild/moderate dysarthria adequately represented by a reasonable number of phonetic-contrast categories?*
- *Is there close agreement between the phonetic-contrast error profiles identified by (a) different listeners and (b) the same listener on different listening occasions?*

- *What is the threshold for dysarthria detection using single-word intelligibility?*
- *Are there phonetic-contrast categories that should be excluded from a clinical assessment of dysarthria because they are equally vulnerable in neurotypical speakers?*
- *Are the number and types of error identified by phonetic-contrast analysis predictive of real-world intelligibility?*

2.1.6. Identifying articulatory errors: Open vs. closed response mode

It was argued in the previous subsection that while it would be *possible* to implement phonetic-contrast analysis using orthographic transcription, a closed response mode would be preferable, especially in a clinical setting, as it would provide a more efficient means of identifying and coding errors. Furthermore, in a forced-choice paradigm, the researcher has complete control over how often each phonetic contrast is tested, meaning that it is possible to calculate the *vulnerability* of a contrast (i.e., the proportion of occasions on which it was not correctly communicated). Contrast categories with higher error rates can then be considered more difficult to produce (leaving aside perceptual distinctiveness arguments). In an open response format, on the other hand, the denominator (the number of times the category was 'tested') is unknown, as it depends on the functional load of the contrast within the speech sample, as well as on other factors, such as the way in which other contrasts in the word are produced. Therefore, in orthographic transcription, a category that is found to yield a large number of errors could signify that the contrast was consistently difficult for the speaker to produce (i.e., that it would yield a large numerator in a closed-response paradigm), but it could also be a consequence of the fact that the category was 'tested' much more often than other categories. Thus, if the goal of the therapist or researcher is to identify phonetic contrasts that represent the greatest production challenge for the speaker, then this is most easily achieved through the use of a closed-response paradigm.[14]

Despite the advantages of the forced-choice approach, it was concluded in Section 2.1.5 that, even if the Dutch intelligibility test can eventually be administered this way, the initial analysis of segmental errors in the present study should be conducted via orthographic transcription, to be certain of capturing the full range of common contrast errors. An open response mode also presents other advantages: (1) It allows all three word segments to be analysed, meaning that there are three times as many opportunities for detecting errors. (2) It avoids the potential sources of bias associated with a multiple-

[14] Note, however, that once an open-response assessment has been administered to a large number of speakers with dysarthria, then the assessor gains an impression of the average vulnerability of each contrast category, and it becomes possible to identify contrasts that present a particular problem for a given individual.

choice paradigm (mentioned in Section 2.1.5), such as forcing the respondent to choose between a consonant error and a vowel error. (3) A frequently-occurring contrast error in an open response format can be considered to be functionally important, irrespective of whether the cause is articulatory difficulty or high functional load. In a multiple-choice response format, on the other hand, additional information about functional load would be required in order to determine the potential importance of a contrast error for real-world intelligibility. A variety of equations and procedures have been proposed to calculate functional load, ranging from simple calculations of the frequency of occurrence of a phoneme within a person's speech (e.g., Pye et al., 1987) to complex models that take into account a wide range of properties of both the target phoneme and the set of phonemes with which it may contrast (e.g., Brown, 1988). As will be shown in Chapter 3, published data on the functional loads of phonemes (and, even more so, phonemic contrasts) are difficult to come by. Furthermore, the studies that do exist often show substantial disagreement. Therefore, a limitation of the closed response format is that it may not be straightforward to use the outcome measures to identify articulatory errors that are important for real-world intelligibility in the sense that they occur frequently in everyday speech.

In light of the above discussion, it was decided that the current thesis would employ an open as well as a closed response format, in a two-stage design. In an initial study, orthographic transcription would be used to determine the feasibility of reducing the speakers' phonemic-substitution errors to a reasonably small set of phonetic-contrast categories. Assuming that such a reduction were possible, a follow-up study would then use a multiple-choice format to allow measurement of the vulnerabilities of these categories. The logical question that arises from such a design is:

> *Are there significant differences between the word-accuracy values and error profiles yielded by an open and a closed listener response format?*

The remainder of this subsection briefly examines the existing evidence in relation to this question for speakers of English (to the best of the author's knowledge, there are no studies that compare the two response modes for speakers of Dutch).

Kent et al. (1989) briefly discussed the issue of the choice of response format when they introduced their intelligibility test, stating only that multiple-choice protocols produce higher scores, which is of "negligible concern" given the relative nature of intelligibility metrics. This opinion was based on two studies by Yorkston and Beukelman (1978; 1980), which showed that multiple-choice intelligibility scores were significantly higher than those elicited from an open-response format, but that both formats ranked the

speakers similarly. As pointed out by Vigouroux and Miller (2007), however, the Yorkston and Beukelman studies have limited generalisability for a number of reasons, including the relatively small sample sizes (between 9 and 12 speakers). Thus Vigouroux and Miller (2007) carried out their own investigation involving 27 speakers with Parkinson's disease and 24 neurotypical control speakers. They used an in-house intelligibility test, tailored for speakers from North East England, which consisted of sixty items (mainly single words plus some short phrases). In the multiple-choice version of the assessment, there were 11 distractors for each target. Word identification was carried out by a pool of 61 naïve listeners such that each speaker was scored by three listeners. As in the Yorkston and Beukelman studies, the closed-format intelligibility scores were found to be significantly higher than the open-format scores (where both were calculated as word accuracy). However, while the correlation between scores for the two formats was significant ($r = 0.721$, $p < 0.001$ for speakers with Parkinson's disease; $r = 0.686$, $p < 0.001$ for control speakers), the relationship between the two scores was sufficiently variable across speakers such that the two methods did not rank them similarly. The authors concluded that open- and closed-format intelligibility tests are qualitatively different.

The finding that a closed response mode yields higher intelligibility scores seems to be intuitive. If nothing else, the opportunity to guess in a multiple-choice test would increase the probability of a correct response. However, this logic applies to the more familiar notion of a 'test' in which there are right and wrong answers and the respondent is not required to interact with the test stimuli. Yet articulatory errors in dysarthria are, in the majority of cases, distortions rather than substitutions (see Section 2.1.3), implying that variables such as word frequency or the listener's response format may influence how such tokens are perceived. Indeed, this was found to be the case in Bunton and Weismer's (2001) paper on vowel-height errors. The authors hypothesised that when vowel tokens are produced in a non-prototypical way (from an acoustic point of view), the response of the listener might differ depending on whether or not there is a near-category option (in their case, a vowel with a different tongue height) to choose from. To test this hypothesis, five experienced phoneticians carried out broad transcription of some of the targets that had been misperceived as vowel-height errors in a multiple-choice assessment with relatively inexperienced listeners (undergraduate students on a communication disorders programme). The phoneticians were blind to the target words and foils. It was found that less than 15% of the transcribed items matched the tongue-height error foil, with the majority of transcriptions agreeing with the intended target. According to Bunton and Weismer (2001), this suggests that when an intelligibility test produces a phonemic error, this is likely to represent an interaction between the extent to which the acoustic

properties of the production deviated from "ideal" and the response format used to detect the error (i.e., free- versus forced-choice). A second possibility is that the discrepancy reflects differences between expert and inexperienced observers in that the latter may have a lower tolerance for phonetic distortions before they perceive a phonemic error. Therefore, it would be preferable for future comparisons of the two response modes to use listeners of equivalent knowledge and experience. Nevertheless, Bunton and Weismer (2001) draw our attention to the possibility that the multiple-choice format could actually *encourage* errors, at least in the case of some contrast categories, by presenting an alternative to the listener that s/he may not have otherwise considered.

To summarise so far, the existing evidence suggests that a closed response format tends to result in higher word accuracy than an open format, although there may be particular circumstances (e.g., specific types of error) for which the forced-choice method actually encourages errors. A second conclusion is that it seems unlikely that the relationship between the intelligibility metrics for the two formats will be entirely consistent across speakers, meaning that a given cohort is likely to be ranked in a different order depending on which format is used. The extent of the disagreement is likely to depend on a number of factors – in particular, the range of severities among the speakers, which determines the gaps between their abilities (Vigouroux & Miller, 2007).

Turning our attention to the second part of the question posed above (i.e., whether there are differences in the *types* of error yielded by the two response formats), based on theoretical arguments, a significant difference seems likely. Firstly, as mentioned, the distractors in a multiple-choice assessment need to be chosen such that the error rate can be measured with reasonable reliability for each of the phonetic-contrast categories, while the distribution of potential contrast errors in orthographic transcription is determined by other factors, particularly the functional loads of the contrasts in the word list. Therefore, contrasts with a high functional load are likely to yield more errors in an open format than in a closed format. Secondly, the aforementioned sources of bias in the multiple-choice format are likely to cause some errors to be supressed and others to be augmented, relative to the open response mode where the listener can report errors on multiple word segments simultaneously. To the best of the author's knowledge, there is only one empirical study that has attempted a qualitative comparison of the two methods (Bunton et al., 2007). In this study, the authors administered a shortened version of Kent et al.'s (1989) assessment, consisting of 53 words, to five adult male speakers with Down syndrome. The multiple-choice responses were provided by ten inexperienced listeners, while the open responses (broad transcription) were obtained from five experts. Unlike

the dysarthria studies mentioned above, the authors found no significant difference in intelligibility scores between the two response modes. However, it is difficult to compare Bunton et al.'s (2007) findings with those of previous studies, as it seems that the intelligibility score for broad transcription in Bunton et al.'s study was phoneme (rather than word) accuracy, although the authors did not make this entirely clear. Furthermore, the use of broad (as opposed to orthographic) transcription could result in a different relationship between the open- and closed-response scores. Regarding the errors themselves, Bunton et al. reported that although the categories that yielded the most errors (pooled across all five speakers) were highly similar in the two response modes, there were some differences; e.g., the 'glottal-null' error was common in the multiple-choice mode but rarely reported in broad transcription. Furthermore, although the *identities* of the top error categories were similar, there were differences in their *rankings*; for example, vowel-duration errors were ranked more highly for the multiple-choice task. The authors also provided some specific examples of differences between the two response modes that may be suggestive of bias. Most notably, listeners preferentially identified vowel errors over consonant errors when both were present in a single word (e.g., *big* identified as the foil *bag* in the multiple-choice mode, but reported as /bæk/ in broad transcription). Thus, it was rare for listeners to transcribe a vowel error when one had not been identified in the multiple-choice task (3.8% of tokens), while 37% of consonant errors identified via broad transcription were not recorded in the closed response mode. Overall, while this study seems to support the hypothesis that there are likely to be significant differences between the qualitative information yielded by the two response modes, it has a number of limitations. In particular, the methods are described very briefly and it is unclear how the authors calculated error rates for the two techniques in such a way that they could be meaningfully compared. As will become evident later in the thesis, such a comparison is by no means straightforward, due to the aforementioned issue of the lack of a denominator in an open response mode. In addition, methodological differences (e.g., in the language, aetiology and method of transcription) mean that Bunton et al.'s (2007) findings may not be applicable to the present research.

2.2. Segmental errors in Dutch speakers with dysarthria

Although the main aim of this thesis is to contribute knowledge of a methodological nature, this knowledge is not without context, and it is possible that the answers to some of the questions posed in Section 2.1 (e.g., whether the range of phonemic errors observed in speakers with mild/moderate dysarthria can be reduced to a finite set of phonetic-contrast categories) will be specific to the language of the population sample. Therefore, it

is worthwhile examining whether there are any data regarding segmental errors in Dutch speakers with dysarthria that could be useful for developing the current set of hypotheses or for informing the study design. The Belgian Dutch phonological system is described in detail in Chapter 3, but briefly, the consonants comprise plosives (/p, b, t, d, k/), fricatives (/f, v, s, z, x, ɣ, h/), nasals (/m, n, ŋ/), liquids (/l, r/) and semivowels (/j, w/). The vowel system consists of 12 monophthongs and 3 diphthongs.

To the best of the author's knowledge, there are no studies that report the types of phonemic or phonological errors typically observed in Dutch speakers with dysarthria as judged by perceptual analysis (either narrow transcription or methods such as phoneme or word intelligibility). Jonkers et al. (2014) asserted that dysarthria in Dutch is "very similar" to English. The source of this statement was not cited, but it may have been based on a combination of clinical experience and theoretical argumentation: the two languages share fairly similar phonologies, including a rich vowel system and complex consonant clusters at both syllable-initial and syllable-final positions. Accordingly, it seems likely that the Kent et al. (1989) categories will capture many of the phonemic errors perceived in Dutch speakers, as these categories were chosen based on a literature review of dysarthric errors in English speakers. Nevertheless, some differences may arise for particular phonemes. For example, in Belgian Dutch, the rhotic is typically produced as an alveolar trill (Verhoeven, 2005), meaning that the contrast categories proposed by Kent et al. (1989) for the English rhotic, /l/-/r/ and /r/-/w/, may not be applicable. A further difference arises in the production of word-initial voiced and voiceless plosives. Whereas English voiceless stops are aspirated in word-initial position, the voiceless plosives in Dutch are unaspirated and the most reliable cue to the voicing distinction (in word-initial position) is the presence or absence of prevoicing (van Alphen & Smits, 2004). It is possible that these different mechanisms for producing the voicing contrast in the two languages could result in a different frequency or type of voicing error. A typical voicing error in Dutch speakers with apraxia of speech, for example, is increased prevoicing in the voiced plosives, a form of hyper-articulation (Jonkers et al., 2014).

There are a small number of studies involving Dutch speakers with dysarthria that analyse acoustic measures related to segmental speech. While the distortions observed in these studies do not necessarily imply that a phonemic substitution would be perceived, they could still be useful in predicting the phonemes and phonetic contrasts that are likely to be most vulnerable. Verkhodanova and Coler (2018) showed that various acoustic measures of vowel centralisation could be used to distinguish between Dutch speakers from the Netherlands with Parkinson's disease (n = 15) and normal controls (n = 15). The

variables in question, which were measured from spontaneous speech, were vowel space area, vowel articulation index, and the F2 ratio of the vowels /i/ and /u/. All of these metrics were found to be lower for speakers with PD, which is indicative of vowel centralisation. Jonkers et al. (2014) argued that vowel centralisation is likely to be a stronger determinant of intelligibility for Dutch than for languages with simpler vowel systems.

Van Nuffelen et al. (2009b) used acoustic features to produce three predictive linear-regression models of subjective phoneme intelligibility as measured by the NSVO (de Bodt et al., 2006). In the first model, the explanatory features were phonemic, in the second they were phonological, while in the third, a combination of both types of feature was used. A five-fold cross-validation paradigm was employed such that in each trial, the models were trained on 80% of the speakers and then tested on the remaining 20%. The sample consisted of 160 pathological speakers (60 dysarthria, 12 children with cleft palate, 42 hearing impairment, 37 laryngectomee, 7 dysphonia and 2 glossectomy) as well as 51 control speakers. Although the models were trained on the full set of speakers, during the testing stage, correlation values between the predicted intelligibility scores and the perceptual intelligibility scores from the NSVO were only calculated for speakers with dysarthria. Since these correlation values were high (0.79 for the phonemic model, 0.83 for the phonological model, and 0.94 for the combined model), it can be concluded that the final set of explanatory features in each model was strongly associated with phoneme intelligibility in speakers with dysarthria. Before discussing the most influential features, it is important to understand how the models were derived. The first step was to transform the utterances (the 50 CVC words from the NSVO) into mel-frequency cepstral coefficients (MFCCs) for successive, overlapping time frames (each of 30 ms duration). The temporal MFCC data were then aligned with the canonical phonemic transcription of the uttered word; i.e., each frame was assigned to one of the phonemes of that transcription. In the case of the phonemic model, neural networks were then used to estimate the posterior probability that, based on the MFCC data, the intended phoneme (according to the transcription) was present in the frame. These probabilities were calculated based on statistical acoustic models of phonemes (including their phonetic contexts) that had been trained on speech samples from a large number of neurotypical speakers. The final output for each speaker was a set of 35 features, where each feature estimated how well, on average, a given phoneme of Dutch was articulated by that speaker based on the acoustic data. These features were used as inputs to the predictive model, and the number of features in the final model was chosen to be safely below the point where the performance started to drop due to overtraining. A similar process was

employed to produce the other two models (phonological and combined), where the set of phonological inputs consisted of 48 features (e.g., 'voicing', 'no voicing', 'velar', 'not velar'). The optimal number of explanatory features in the final models was 15 for the phonemic and phonological models, and 34 for the combined model. Table 2.5 presents these features for the first two models only (the results are similar for the combined model). It is important to appreciate that the listed features are those that showed an *association* with overall intelligibility of sufficient magnitude to contribute to the regression model. Therefore it is possible that some of the features in Table 2.5 did not, on average, correspond to poorly articulated phonemes and phonological features (i.e., a speech feature can be strongly associated with intelligibility without yielding high average error rates; see Section 2.3).

Phonemic model	*Phonological model*
/ɔu/	Lateral
/l/	Low †
/ø/	Mid-high †
/#/	High †
/ɑ/	Not mid-high †
/ɔ/	Not mid-low †
/z/	Not high †
/s/	No silence
/ɛ/	Not low †
/x/	Velar
/t/	Closure *
/ɪ/	Fricative
/j/	Not velar
/p/	Mid ‡
/ɣ/	Silence

* Contact between articulators when producing a plosive

† Refers to the vertical tongue position

‡ Refers to the horizontal tongue position

Table 2.5. The most important phonemes and phonological features for predicting overall phoneme intelligibility in Dutch speakers (160 pathological, 51 control); van Nuffelen et al. (2009b). The authors explain that since silence has no pathological meaning, this feature models the acoustic background, which is subtracted from the other features during regression.

Table 2.5 reveals that many of the phonemic and phonological features relate to vowels, and seven of the phonological features involve vowel height. This implies that the extent of a speaker's vowel-height distortions is strongly associated with their overall phoneme intelligibility. Secondly, the phoneme /l/ and the corresponding phonological feature 'lateral' are important predictors of the perceptual intelligibility measure. The authors state that no substantial evidence for the importance of these features can be found in the literature, although they note that the Kent et al. (1989) test includes an /r/-/l/ substitution. Thirdly, it can be seen that four of the six Dutch fricatives (the other two being /f/ and /v/)[15] are important features of the phonemic model, while the feature 'fricative' is included in the phonological model. The implication is that accurate production of this manner of articulation is correlated with overall intelligibility. The remaining phonemes listed in Table 2.5 are /j/, /t/ and /p/. The importance of the phonological feature 'contact' suggests that the acoustic distortions for the two plosives may, at least in part, be a consequence of incomplete closure. Finally, the velar place of articulation (both its presence and its absence) is predictive of overall intelligibility. Van Nuffelen et al. (2009b) state that the literature does not seem to reveal any evidence that motivates the selection of the velar-related features. However, this could perhaps be a consequence of the fact that English (the language investigated in most previous studies) does not have a fricative at the velar position; that is, the distortions that encompass these features in van Nuffelen et al.'s data might pertain to fricatives. Indeed, the voiced and the voiceless alveolar and velar fricatives (/z, s, ɣ, x/) all appear within the list of the most important phonemes. If errors in tongue advancement were to occur for these sounds, then some of those errors would have relevance to the phonological features 'velar' and 'not velar'. Finally, it is interesting to note that neither voicing nor the lack of voicing is an important phonological feature. This implies either that the voice contrast was reasonably robust or that distortions related to voice were relatively consistent across speakers of different intelligibility levels.

To summarise, there is a lack of direct information about the types of phonemic substitution errors seen in Dutch speakers with dysarthria. There is some anecdotal evidence, as well as evidence from acoustic studies, to suggest that there will be significant overlap with the articulatory errors seen in English speakers. However, there are also likely to be differences, particularly because there are phonemes of Dutch that are not part of English phonology and that can be considered "difficult" to produce, such as

[15] Dutch phonology also includes the voiced glottal fricative, /h/; yet van Nuffelen et al. (2009b) only refer to six fricatives. The reason for this is unknown, but it could reflect the fact that in production terms, /h/ is closer to a vowel than a fricative.

the rhotic (an alveolar trill in the Antwerp accent) and the velar fricatives. The frequencies and types of substitution error seen for these phonemes could turn out to be important markers for dysarthria, much like the alveolar-palatal fricative contrast in English. This leads to the final research question provoked by this literature review:

> *What are the phonemic and phonetic-contrast errors of Belgian Dutch speakers with dysarthria?*

2.3. Relationship between segmental speech errors and overall intelligibility

As explained in Chapter 1, this thesis considers a scenario in which the therapist and client anticipate the potential value of identifying the segmental, articulatory errors that arise during the utterance of single-word stimuli. It is possible to imagine situations in which information of this nature would almost certainly be useful; for example, for the purposes of programming an AAC (augmentative and alternative communication) device based on single-word commands, or for facilitating communication with an individual who has co-occurring non-fluent aphasia and is only able to communicate using single words or very short phrases. However, in many cases where articulatory analysis is conducted, single-word utterances are not the primary means of communication; rather, the underlying assumption is that the errors observed in a single-word production task are also a major cause of reduced intelligibility in spontaneous (or other types of connected) speech. However, this assumption has not been extensively tested. Furthermore, it is relatively easy to construct arguments as to why one might not expect a strong correlation between intelligibility in single-word reading and intelligibility in connected speech. Firstly, the two sets of speech stimuli may differ with respect to phonetic characteristics (e.g., the distribution of phonemes or phonetic contrasts). Secondly, the processes involved in articulating single words differ from those that underlie the production of connected speech, which could result in differences in the types of segmental error *produced* in the two tasks. Thirdly, from a *perceptual* perspective, the additional cues available in connected speech (e.g., semantic, prosodic) are unlikely to confer equal benefit (or equal disadvantage in the case where such cues are missing or distorted) on all phonetic contrasts. The remainder of this section is devoted to examining the research literature on methods of analysing the relationship between specific speech errors and overall measures of intelligibility. To provide a broader understanding, the review begins with a description of studies in which the overall intelligibility metric was derived from single words rather than from connected speech.

Many of the previous studies that implemented phonetic-contrast analysis also examined the link between the error rates for specific contrast categories and overall intelligibility (word accuracy). Some of these studies used relatively simple methodologies, such as a visual comparison of the contrast-error profiles for different groups of speaker defined by overall intelligibility (e.g., Kent et al., 1989), or calculation of the correlation coefficient between word-intelligibility scores and the mean error rate across several of the most affected categories (Kent et al., 1990). Other studies used stepwise multiple-regression to determine which of the phonetic-contrast features were most predictive of word intelligibility (e.g., Weismer & Martin, 1992; Whitehill & Ciocca, 2000b). Due to the fact that error rates for different contrast categories tend to be inter-correlated, it is often possible to achieve reasonable predictive power using just a few categories. For example, in their analysis of 25 male speakers with ALS, Weismer and Martin (1992) showed that two categories ('stop vs. nasal' and 'initial /h/ vs. null') together accounted for over 95% of the variance in word-intelligibility scores. Obviously, the high predictive power in these studies is also a consequence of the fact that the same data were used to calculate the individual error rates and the overall intelligibility measure.

The results of such correlation and regression analyses may have important applications. For example, they could be used to develop an efficient means of intelligibility testing, based on a small number of highly sensitive variables (Kent, 1992). They could also be used to aid understanding of the predominant underlying speech deficits; e.g., Weismer and Martin's (1992) findings about the predictive power of the stop-nasal and glottal-null contrasts are consistent with the belief that ALS involves impairment to the velopharyngeal and laryngeal subsystems of speech. However, the outcomes of such analyses do not enable the selection of targets for therapy because:

(1) Functional load is not taken into account. By way of illustration, Weismer and Martin (1992) found that 82% of the variance in single-word intelligibility scores could be accounted for by the 'alveolar vs. palatal fricative' contrast in female speakers with ALS. Yet this contrast has a relatively low functional load in the English language and is likely to have only a small effect on real-world intelligibility.
(2) High correlation does not imply high vulnerability. It is theoretically possible for an error to be a powerful predictor of single-word intelligibility while in fact producing relatively few errors (on average) in the studied population. This was found to be the case, for example, for fricative-affricate confusions in Whitehill and Ciocca's (2000b) study of Cantonese speakers with cerebral palsy.
(3) Substitution errors detected in single-word reading may not be a barrier to intelligibility in connected speech.

The remainder of this subsection reviews the state of knowledge about the relationship between articulatory errors and intelligibility in connected speech. One of the contributions of Darley et al.'s (1969a) seminal study was to provide correlation coefficients between the mean rating values for their 36 individual perceptual dimensions and the mean rating value for the overall dimension referred to as "intelligibility". As was also the case for the individual dimensions, intelligibility was rated on a 7-point scale. In most cases, the connected-speech sample was the Grandfather passage, although in some instances, the speakers provided a conversational sample, and a very small number performed a sentence-repetition task. The correlation values were calculated separately for each of the seven neurological groups. The dimension "imprecise consonants" was among the top five most deviant speech characteristics for all neurological groups and was consistently highly correlated with intelligibility (range 0.77 – 0.92).

De Bodt et al. (2002) conducted a study to determine the extent to which each of four individual speech dimensions (voice quality, articulation, nasality and prosody) contributes to the overall intelligibility of connected speech. The speech data were obtained from Aronson's (1993) recordings of English speakers with dysarthria. The speakers (n = 79) were selected from the database at random, but it was ensured that they represented a variety of dysarthria types. Two experienced listeners assessed the speech samples (either the Grandfather passage or spontaneous speech) by rating each of the aforementioned speech dimensions, as well as the overall intelligibility, on a four-point scale. Firstly, the authors calculated correlation coefficients between the individual dimensions and intelligibility, which were found to be 0.82 for articulation, 0.46 for voice quality, 0.32 for nasality, and 0.55 for prosody. They then computed a multiple linear regression model (R^2 = 0.90), which yielded regression coefficients of 0.66 for articulation, 0.16 for voice, 0.01 for nasality, and 0.31 for prosody. Finally, the model was tested on a group of 16 Dutch speakers with dysarthria. The intelligibility ratings of the listeners were compared with intelligibility values calculated from the model (based on listener ratings of the individual dimensions). For 12 of the 16 test speakers, the judged and calculated intelligibility scores were in agreement (within the 95% confidence interval of the calculated values). The main insight provided by this study, from the current perspective, is that articulation appears to be the most important speech dimension for connected-speech intelligibility.

A drawback of de Bodt et al.'s (2002) study is its reliance on subjective perceptual ratings, which may be prone to bias. Rong et al. (2016) suggested an intelligibility model based on physiological indices, which the authors state were chosen in preference to acoustic

indices (e.g., Lee et al., 2014) because the former can unambiguously represent the status of individual speech subsystems. Rong et al. (2016) investigated 66 speakers with ALS and obtained a variety of instrumental measures to quantify the changes in four subsystems of speech (articulation, resonance, phonation and respiration) over time. A sentence intelligibility test was performed to obtain an overall measure of intelligibility (the percentage of correct words). The model itself was highly intricate and included a variety of preliminary stages to reduce the data (i.e., number of predictors), replace missing values, and check for potential confounding factors. Furthermore, since the trajectory of the decline in intelligibility in ALS is thought to bi-phasic, the authors compared linear models with bi-phasic nonlinear models to determine the best fit. They also used a mixed-effects paradigm to account for the heterogeneity among individuals. The stepwise regression model with the best performance consisted of five predictors, which together accounted for 95.6% of the variance in the overall decline of intelligibility. The articulatory subsystem showed the greatest contribution (57.7%), followed by the resonatory subsystem (22.7%), the phonatory subsystem (8.3%), and the respiratory subsystem (7.2%). The two prominent features identified for the articulatory subsystem were slowed lip and jaw movements and slowed AMRs (alternating motion rates).

A number of studies have examined the relationship between specific articulatory errors and an overall measure of intelligibility derived from connected speech. Coppens-Hofman et al. (2016) identified phonological errors in single words (obtained from picture naming) and used a 5-point rating scale to represent intelligibility in spontaneous speech. The phonological error categories were fairly general, such as the proportion of correct consonants in syllable-initial position and the proportion of correct syllable-initial consonant clusters. The population consisted of 34 Dutch-speaking adults with learning disability of mixed aetiology. The authors performed separate regression analyses for two groups (mild vs. moderate learning disability, defined by IQ). In the mild group, four error categories together predicted 79% of the variance in SSI, while three error categories predicted 69% of the variance in the moderate group.

Kuruvilla-Dugdale et al. (2018) developed a scheme to calculate a metric of articulatory precision based on narrow transcription. The method was applied to 16 multisyllabic words (e.g., *phantom*, *hospitable*) uttered by American English speakers (n = 8) with dysarthria due to ALS. If all the phonemes in a word were produced accurately, then the accuracy score was 1. Otherwise, for each phoneme that was misarticulated, its baseline accuracy score (equal to 1 divided by the number of phonemes in the word) was degraded using a weighting factor based on the contribution of the misarticulation to

speech intelligibility, as judged from the literature. Specifically, they chose the following weighting factors: additions reduced the phoneme accuracy score by 25%, distortions by 50%, substitutions by 75%, and omissions by 100%. As an example of their reasoning, omissions were weighted most heavily because (i) they are significant predictors of intelligibility decline in dysarthria and (ii) they result in a change in the syllable and word structure. The metric of articulatory precision was able to separate control speakers from a subset of the speakers with dysarthria (n = 4) who were classified as moderate-severe, where severity was assessed using word accuracy in the sentence intelligibility test (SIT; Yorkston et al., 2007). Specifically, the moderate-severe ALS group had mean precision scores that were two standard deviations below the mean of the control group. There was no significant difference in articulatory precision between control speakers and the mildly impaired group (n = 4), and in fact, the mean precision of both groups was close to ceiling (100%).

The final study worth mentioning in this genre is that of Wilson et al. (2019). This paper formed part of a large research project that aimed to assess the relationship between intelligibility in adolescents with Down syndrome (n = 45) and a wide variety of variables including, but not limited to, speech variables. The SSI metric was calculated by dividing the number of intelligible words in a conversational speech sample by the total number of words produced. The result was then converted to a three-point ordinal scale ('low', 'medium' or 'high') to account for the fact that the scores were not normally distributed. The most notable methodological feature of Wilson et al.'s (2019) study is that the segmental deficits (which were identified using narrow transcription) were derived from the same conversational speech samples as used to measure SSI. The authors observed that the 'low' intelligibility group was characterised by across-the-board deficits in most phonemes and phonetic features, especially for vowels, whereas speakers in the 'high' intelligibility group showed a more selective pattern of deficits. They claimed that the broad spectrum of errors in low-intelligibility speakers differs from the pattern shown in other studies of people with Down syndrome, which have reported prototypical error classes. Wilson et al. attributed this difference to the fact that previous studies assessed intelligibility (and presumably also articulatory errors) using single-word tasks.

The fact that Wilson et al. (2019) derived segmental errors from a spontaneous-speech sample ought to improve the strength of the relationship between articulatory errors and intelligibility, relative to studies in which articulatory errors were identified in single words. However, it remains the case that the findings of such studies do not imply *causation*. Whether the observation is a higher error rate for a particular feature in

speakers who are less intelligible or a strong correlation between a specific error rate and overall intelligibility, in neither case would it be legitimate to conclude that the error itself reduced the accuracy with which the message was perceived. It could be the case that the error in question co-varied with another feature that lowers intelligibility. Thus the situation is best summarised by Weismer et al. (2001), who noted that while a variable may explain variation in intelligibility across speakers, it may not be an "integral component" of intelligibility in the sense that it would yield an improvement in intelligibility following treatment. To identify variables that are integral components of real-world intelligibility, an *explanatory* approach is needed rather than one based on association. For example, one could measure the degree of improvement in connected-speech intelligibility following either targeted intervention (e.g., Tjaden et al., 2014; Beijer et al., 2014) or computer processing of the speech signal (e.g., Rudzicz, 2011). Alternatively, one could examine the effect of simulated speech errors, produced either by a speech synthesiser (e.g., Rudzicz, 2011) or an unimpaired human speaker (Klein & Flint, 2006), on connected-speech intelligibility. A further possibility would be to study across-utterance variations in intelligibility within a given speaker. As discussed by Yunusova et al. (2005), the identification of variables that are correlated with *intra*-speaker intelligibility is particularly important because such variables are likely to be under the control of the speaker (and, hence, amenable to therapy). Yunusova et al. (2005) studied linguistic variables (e.g., utterance duration) that were expected to be reasonably independent of the phonetic composition of the utterances. This design could be implemented in the other direction: that is, studying the correlation values between *phonetic* variables and the overall intelligibility of the utterance, where the variables in question are chosen such that they are likely to be independent of linguistic factors. Examples of suitable phonetic variables might include speech rate and pitch variation. However, it would not be straightforward to employ this technique to study the impact of a specific segmental feature, such as a phonetic-contrast category, as this would require the careful design of connected-speech stimuli that differ only with respect to specific speech sounds. A large number of such utterances would then be required in order to study a wide range of phonetic-contrast categories. In addition, the speakers might need to be trained to ensure that they produce the utterances in a consistent way with regard to all other linguistic and phonetic features such as pitch, loudness and speech rate.

The conclusion of the above discussion is that the implementation of an explanatory approach would be a significant undertaking, one that would be beyond the scope of the present thesis. Nevertheless, the assumption that *articulatory errors play an important role in real-world intelligibility* is central to the thesis. Therefore, it was considered

important to contribute to the evidence base for this assumption to an extent that was possible within the available resources – in particular, the limited time frame of the project and the relatively low number of speakers with dysarthria (10). It was decided that an investigation would be carried out to examine the degree of correlation between an intelligibility measure derived from single-word reading and an intelligibility measure derived from natural, unconstrained spontaneous speech. It is important to appreciate that the goal of this investigation was relatively modest – to gain a preliminary indication of the level of correlation. A thorough examination of the relationship between single-word reading intelligibility and spontaneous-speech intelligibility, including a rigorous investigation of potential confounding factors (e.g., coherence, speech rate, fluency and prosody) would have been a significant undertaking and was beyond the scope of this thesis, the main purpose of which was to improve understanding of the methodology for identifying *segmental* errors. Therefore, the goal was clinically motivated: to assess the importance of articulatory errors for intelligibility in spontaneous speech. A high correlation coefficient would indicate, at the very least, that substitution errors in a single-word reading task co-vary with the factors that affect intelligibility in spontaneous speech. However, further research would then be required to uncover the precise nature of this relationship, including whether substitution errors are a major *cause* of reduced intelligibility in spontaneous speech, and if so, which specific contrasts exert the most influence. If a strong correlation is *not* observed, then this would be a valuable finding for clinicians and theoreticians alike. Future research might then focus on attempting to understand the confounding factors. This would be of considerable theoretical interest, while from a clinical perspective, it would enable identification of the subset of speakers for whom articulatory therapy is expected be most beneficial.

A number of studies have investigated the correlation between single-word intelligibility and a measure of intelligibility derived from connected speech (e.g., Lagerberg et al., 2014; Yorkston & Beukelman, 1978; Yunusova et al., 2005). These studies reported reasonably high correlations (≥ 0.8), which is in agreement with the studies mentioned above that computed the correlation between a subjective rating of articulatory precision and intelligibility in connected speech (Darley et al., 1969a; de Bodt et al., 2002). However, it is difficult to extrapolate from these findings to predict the level of correlation expected in the present study, as the relationship between the two intelligibility measures is likely to depend on a large number of factors, including the language, the distribution of dysarthria severities and types, and the characteristics of the speech stimuli. In particular, studies that used sentence reading as the connected-speech task (Yunusova et al., 2005; Yorkston & Beukelman, 1978) are likely to yield higher correlations than studies (such as

the present one) in which the speakers deliver a monologue, as the latter introduces additional sources of variability between speakers. From a methodological perspective, the study that is most closely related to the current work is that of Lagerberg et al. (2014), as their methods of eliciting spontaneous speech and measuring its intelligibility were chosen to be implemented in the present study (see Section 2.4). Lagerberg et al. reported a Pearson's correlation coefficient of 0.79 ($p < 0.01$) for the relationship between the percentage of correct consonants in a single-word intelligibility test and their measure of spontaneous-speech intelligibility. However, it would not be justified to expect a similar outcome for the present study, as there were substantial differences in various aspects of the study design. Most notably, Lagerberg et al. (2014) investigated an entirely different clinical population, namely Swedish-speaking children (mean age 6 y; range 4;6 – 8;3) with a speech-sound disorder. Setting aside the differences in language and aetiology, the linguistic features of the spontaneous speech of these two groups (i.e., children and adults) are likely to show large discrepancies. For example, differences would be expected in terms of utterance length, lexical and grammatical complexity, and discourse coherence. Furthermore, the two groups are likely to differ in terms of the extent to which they employ compensatory mechanisms, such as a reduced speech rate or greater effort.[16] When hypothesising about the differences between the two groups, it should be borne in mind that the relevant metric is the degree of *variability* in the confounding variable among each set of speakers. Thus, a lower correlation would be expected for adults if, for example, they show greater variation in their degree of prosodic impairment than children, not necessarily because they simply have a higher level of prosodic dysfunction. This makes the task of extrapolating from one study to another even more difficult.

Most of the studies that have measured correlation coefficients between articulatory features and intelligibility in connected speech have assumed that the underlying relationship is linear (e.g., through the use of Pearson's r or linear regression). In fact, there is evidence to suggest that the relationship between single-word intelligibility and connected-speech intelligibility may be nonlinear. Sentence-production tasks tend to increase the intelligibility of speakers with mild dysarthria (relative to single-word reading), but speakers with severe-to-profound dysarthria show a varied trend and may have higher, equivalent, or lower intelligibility in sentences than in single words (Allison et al., 2019). There may be a number of reasons as to why the additional syntactic and semantic cues in sentences can be less beneficial in the case of severe speakers. Firstly, it

[16] The direction of this difference is not known. The children in Lagerberg et al.'s (2014) study were recruited through SLTs, so it is likely that at least some of them were receiving treatment for their speech-sound disorder, which may have included compensatory mechanisms.

is possible that the extra burden of having to produce connected speech, combined with assimilatory and reductive processes, degrades the signal to such an extent that there is no longer any contextual advantage. An alternative explanation can be imagined based on Lindblom's (1990) theory of hypo- and hyper-articulation. Speakers might be less likely to hyper-articulate in connected speech, as they are aware that the listener has the benefit of contextual cues. A lower level of articulatory effort could be more detrimental to the intelligibility of speakers who have a greater level of inherent impairment. In addition to the phenomenon just discussed, another type of non-linearity could arise from a ceiling effect. In other words, if the presence of context results in improved intelligibility, then this improvement will level off as the maximum intelligibility is reached. Such an effect has been predicted by probabilistic models of the influence of context on spoken language (Boothroyd, 2002). If *both* of the aforementioned sources of non-linearity are observed, then this would result in an s-shaped curve, with speakers in the mid-range showing the greatest level of improvement between the two speech tasks. In order detect such non-linear behaviour, a larger sample size than that available in the present study would be needed. Nevertheless, it is worthwhile bearing in mind that nonlinearity might be another cause of reduced correlation when using statistical methods that assume otherwise.

2.4. Quantitative measures of spontaneous-speech intelligibility

The final component of this literature review consists of a brief survey of the methods for measuring intelligibility in spontaneous speech. The decision to use spontaneous speech rather than a reading passage was based on the fact that the primary goal was to test the assumption that articulatory errors play an important role in *real-world* intelligibility, and spontaneous speech is generally considered to have the greatest level of ecological validity. Furthermore, spontaneous speech has received less attention than reading tasks in previous literature (see Section 2.3). It may be the case that spontaneous speech with a *familiar* person is the most common form of communication for many individuals with dysarthria. However, conveying information to an unfamiliar conversation partner, which was the scenario in the present study, is relevant to a wide range of everyday situations, such as social interactions with acquaintances or in public spaces, and conversations with care-workers in an acute or novel setting. Another option would have been to use a task that is constrained to some extent (often referred to as 'semi-spontaneous speech'), such as picture description, a request for procedural information, or concurrent commenting – a method in which the speaker provides a commentary on the events in a silent video that the listener has also viewed (Alves et al., 2020). However, such tasks usually involve some degree of prior knowledge of the target words and/or narrative, especially if each listener

assesses multiple speakers, and this knowledge might confer unequal benefit on speakers of different intelligibilities. Furthermore, it was reasoned that in the present population, instead of creating a level playing field, semi-spontaneous speech might even *increase* the variability among speakers with regards to the relationship between the two types of intelligibility measure. This is because the population sample included individuals who had some degree of cognitive difficulties, including executive dysfunction, or were experiencing symptoms such as distress, fatigue and anxiety (the sample included individuals receiving inpatient hospital care). It is possible that participants faced with these sorts of challenges would have had greater difficulty in producing a semi-spontaneous discourse, in terms of factors such as topic maintenance, event sequencing, cohesion and fluency, than in speaking in an unconstrained manner. For similar reasons, it was decided that the speakers would be free to choose the topic of their discourse. In comparison with fixing the topic, this strategy is likely to reduce the cognitive burden on speakers experiencing some of the aforementioned difficulties, such that their intelligibility is not unduly affected by a lack of coherence and/or fluency. In the case of speakers with executive dysfunction, it also avoids the situation where a speaker might be reluctant to speak on a particular topic or is unable to maintain focus on a topic that was not of their choosing.

Another sense in which the speech production task was relatively unconstrained was that no attempt was made to invoke compensatory strategies, such as a reduced speech rate, clear speech or loud speech (e.g., van Nuffelen et al., 2009a; Tjaden et al., 2014). It is possible that the use of such strategies would have increased the level of correlation, by controlling for some of the confounding variables. However, compensatory strategies were considered beyond the scope of this thesis (which focused on segmental errors) and would have added time to the interviews, due to the need to train participants in the appropriate technique. Furthermore, it was not self-evident that the use of compensatory strategies would have rendered the findings easier to interpret. For example, while the rationale behind speech-rate reduction is that it ought to increase articulatory precision, studies concerning the effect of rate control on intelligibility have produced mixed findings. In fact, van Nuffelen et al. (2009a), who recruited 19 Dutch speakers with dysarthria of various types, found that a reduced speech rate did not result in higher mean intelligibility ratings, with significant *decreases* in intelligibility (defined as a reduction of more than 8%) observed in every subject, depending on the method of rate control. One of the explanations offered by the authors for their unexpected findings was that rate-control methods require a certain degree of attention and coordination, which only some speakers are able to achieve.

In light of the above discussion, it was decided that the goal of this part of the study would be to determine the degree of correlation between single-word reading accuracy and an intelligibility metric derived from a natural, uncontrolled form of oral communication (a monologue). If a strong correlation between single-word intelligibility and spontaneous-speech intelligibility is *not* observed, then the finding would still constitute a significant contribution (especially for clinicians), as it would imply that articulatory therapy is not likely to improve the SSI of all speakers, at least not when used as an isolated treatment. It would then be a matter for future research to delve into the confounding factors in order to (a) identify the subset of speakers for whom articulatory therapy is expected to be most beneficial and (b) assess the potential value of combining articulatory therapy with other treatments or strategies.

The most common perceptual[17] approach for assessing spontaneous-speech intelligibility in clinical practice is for the assessor to provide a subjective rating on an ordinal (e.g., a five- or seven-point) scale, which is effectively an equal-appearing interval (EAI) scale. However, it has long been known that the perception of some sensory dimensions is poorly represented by the use of an EAI scale, as respondents tend to exhibit a systematic bias towards subdividing the lower end of the continuum into smaller intervals (Zraick & Liss, 2000). Indeed, intelligibility appears to be an example of a perceptual dimension that is not perceived linearly (Schiavetti, 1992). Furthermore, due to the prescribed nature of an EAI scale, it may not capture the respondent's full range of perception (Zraick & Liss, 2000). For these reasons, research studies that call for a subjective rating of intelligibility often favour a method known as DME (direct magnitude estimation), which is based on ratio scaling. DME does not make linear assumptions about the perception of the dimension in question and it is not bound by fixed minimum or maximum values. The preferred method of implementing DME involves comparing the speech sample with a constant reference sample chosen to represent the midrange intelligibility level (Weismer & Laures, 2002). It is convenient to assign a numerical value to the reference sample (e.g., 100) and then score other stimuli relative to that standard. Despite its advantages over the EAI method, DME still suffers from some significant disadvantages. In particular, it remains a subjective procedure that is prone to bias from a variety of sources, including different internal standards and varied experience. This made it unsuitable for the current project where there was considerable variation among listeners in terms of their level of experience and training, as well as their listening environment (see Chapter 3). In

[17] Automated methods of calculating intelligibility based on acoustic analysis were not considered for implementation, as this would have required additional research and development in an area that was beyond the scope of the thesis.

addition, Weismer and Laures (2002) reported that a set of sentence-level utterances, obtained from four individuals with dysarthria and three neurotypical speakers, was scaled differently depending on the identity of the standard. Finally, the assessment of speech using DME requires either experienced SLTs or individuals who have received a reasonable amount of training (Zraick & Liss, 2000).

Another alternative to EAI scales is the visual analogue scale, VAS (e.g., Tjaden et al., 2014; Stipancic et al., 2016; Fletcher et al., 2017) in which the listener is presented with a continuous horizontal or vertical scale for subjective magnitude estimation. Each endpoint carries a description of one of the extremes of the statement to be evaluated (e.g., "cannot understand anything", "understand everything"). The listener then rates the level of intelligibility by using the mouse (in the case of a computerised VAS) to place a marker at the desired position along the line. This results in an output on a scale between, say, 0 and 100. Unlike equal interval scales, the VAS method does not force listeners to partition their judgment of intelligibility into categories, and it may enhance listener ability to index differences in speakers' intelligibility levels (Fletcher et al., 2017). In their study of New Zealand speakers with dysarthria due to a variety of aetiologies, Fletcher et al. (2017) used a VAS to obtain ratings of both ease of understanding (i.e., intelligibility) and articulatory precision based on a reading task (the Grandfather passage). For each listener, the raw VAS scores were converted to *z*-scores based on the mean and standard deviation of all the ratings provided by that listener. The authors found that the VAS method enabled listeners to record their judgments quickly, and that the raw scores showed high inter- and intra-rater reliabilities. They also reported that: (i) the inter-rater reliability, expressed in terms of intraclass correlation coefficients, was higher for articulatory precision ratings (0.84) than for intelligibility ratings (0.68); (ii) although the precision ratings and the intelligibility ratings were highly correlated, they showed a slight curvilinear relationship. Specifically, for the 'above average' scores, there was less variation among the intelligibility scores than among the precision scores, implying that ratings of intelligibility are not as sensitive to mild dysarthria as ratings of speech precision; (iii) ratings of speech precision were better able to separate the speakers with dysarthria from healthy controls; and (iv) the speech-precision ratings were able to explain a greater amount of variance in metrics of vowel dispersion than the intelligibility ratings. The authors concluded that the VAS method shows promise for producing reliable perceptual ratings that are strongly associated with instrumental measures, especially when careful consideration is given to the rating variable (e.g., precision instead of intelligibility). Stipancic et al. (2016) likewise showed evidence for the utility of the VAS method by comparing perceptual judgments of speech severity using a computerised VAS

with accuracy values for key words, where the latter were calculated from orthographic transcription. The listeners had little or no experience of disordered speech. Intra- and inter-listener reliabilities were slightly higher for the VAS task than for orthographic transcription, despite the fact that the former was expected to be more subjective in nature. For each of the 78 speakers, a correlation between VAS scores and transcription scores was computed across all utterances. These correlations ranged from 0.08 to 0.87, with an average of 0.57 (SD = 0.18). It is important to point out that Stipancic et al.'s (2016) study had a number of limitations, including the facts that (i) it only investigated speakers with high intelligibility and (ii) different speech samples were used for the two tasks (the Grandfather passage for VAS and Harvard sentences for transcription).

While the above findings are encouraging, they do not convincingly demonstrate that there is a strong correlation between VAS ratings and a measure of intelligibility derived from orthographic transcription – at least not for every speaker. Furthermore, the degree of correlation has yet to be tested for spontaneous speech. More critically, at the time of designing the studies in this thesis, the aforementioned papers (Stipancic et al., 2016; Fletcher et al., 2017) had not yet been published. Therefore, the VAS method was ruled out for the same reason as other rating procedures: it is ultimately a subjective approach that was thought to be unsuitable for a heterogeneous set of listeners. It was decided that in contrast to these scaling procedures, the current thesis would implement a technique based on *transcription*, which would then be used to calculate a measure that reflects the proportion of correctly perceived speech units (e.g., consonants, syllables or words). As stated, it was expected that such an approach would result in a measure that is more objective and repeatable than a subjective rating. It was also thought that a transcription approach would be more suitable for listeners who have little or no experience in assessing disordered speech.

The main drawback of using transcription lies in the difficulty of knowing exactly what the speaker said, which makes it problematic to calculate an accuracy metric. A possible solution is to work with the speaker to produce a transcript of the spontaneous-speech sample, which is then held to be correct. However, the production of such transcripts would be a time-consuming and challenging undertaking, particularly for some speakers (e.g., those with severe dysarthria, cognitive difficulties, and/or an impaired ability to write or type). Therefore, it was not a practical possibility for the present study, where many of the speakers were only available for a brief period of time and had co-occurring conditions that would have made it difficult for them to produce an accurate transcript. Recently, Lagerberg et al. (2014) proposed a quantitative measure of SSI that was

designed to overcome these challenges. The clinical population consisted of Swedish children with a speech-sound disorder (n = 10), and the listeners (n = 20) were students of an SLT programme (n = 18) plus two recent graduates. Listeners were instructed to transcribe orthographically all the words that they understood and, for the remaining words, to record the number of perceived syllables. Guesswork was discouraged. Intelligibility was calculated as the number of syllables in the transcribed words divided by the total number of perceived syllables. This ratio was found to be highly correlated (r = 0.79) with the percentage of correct consonants from a single-word utterance test. Lagerberg et al. (2014) further reported that inter-judge reliability was excellent when scores represented the average judgment of four different listeners. This suggests that the method would be reliable for use in a research context where it is often possible to use multiple judges. On the other hand, a measure of reliability that was calculated across all listeners on an *individual* basis was found to be fair to poor, with two of the listeners producing very low intelligibility scores (reportedly because they had a reluctance to transcribe words unless they were very certain that they had understood them correctly). The implication is that the technique may not be suitable for clinical work, where usually only one judge is available, although it should be noted that the individual inter-rater reliability values may have been higher for experienced SLTs and/or listeners who had received a greater amount of practice and training. Intra-judge reliability was investigated based on six samples presented to four listeners on a second listening occasion, which yielded a Pearson's correlation of 0.94 (p < 0.01). The score was higher for the second transcription in 75% of cases, which is a common finding when transcribing spontaneous speech. However, in 83% of cases, the difference in SSI values between the two transcriptions was less than 10 percentage points. Finally, for a subset of the cohort (6 speakers), the authors measured the level of agreement across different monologues produced by the same speaker. They reported strong consistency when SSI was measured at the listener-group level: in 72% of cases, the difference in SSI between the two speech samples was less than 10 percentage points. Nevertheless, the authors point out that this source of random variation is a potential drawback of the technique, and that in serial assessment, care needs to be taken to distinguish between changes in intelligibility that are due to treatment and changes that are related to differences between the two monologues (e.g., in terms of content or coherence). In summary, Lagerberg et al. (2014) concluded that their method for assessing intelligibility on the basis of the percentage of syllables perceived as understood has high validity and reliability provided the mean SSI across several listeners is used, which is often possible in a research context. Regarding the clinical applicability, the authors suggested that the method could be used in

situations where the same listener makes both the initial and the follow-up assessment, e.g., when a given therapist evaluates the effect of intervention.

To the best of the author's knowledge, Lagerberg et al.'s (2014) technique has not been tested in adult speakers with dysarthria. However, as argued above, the alternative approaches for measuring SSI were considered unsuitable for the present study, either because they would require too great a time commitment on the part of the speakers (to produce a transcript) or because they would demand a large cohort of expert or highly trained listeners. Furthermore, the fact that Lagerberg et al.'s technique has not been tested in adult speakers could actually be considered as a desirable quality, as gathering evidence regarding the potential usefulness of the technique increases the contribution of this part of the thesis. Given that the thesis, in general, focuses on methodological issues, the goal of examining the applicability of Lagerberg et al.'s method to a new population is in keeping with this ethos. The findings with regard to this aim are likely to have broader relevance for understanding the challenges associated with the quantification of SSI, particularly in cases where a transcript of the speech sample is not available.

2.5. Objectives and hypotheses

The main goal of this thesis, expressed in broad terms, was to improve understanding of the methodological factors that affect the perceptual identification of articulatory errors in Belgian Dutch speakers with acquired dysarthria. The first conclusion reached in the above literature review was that, given the resources available for this project, articulatory errors would be identified using an approach that would be suitable for naïve listeners (either orthographic transcription or multiple choice). As argued in Section 2.1.5, the remaining methodological choices were then somewhat limited. The most logical choice of stimulus was a set of real, monosyllabic, single words that are highly contrastive. Such stimuli (a) allow the articulatory dimension to be investigated in isolation, without interference from other speech dimensions or from linguistic cues, and (b) increase the likelihood that substitution errors will be perceived even in speakers with mild dysarthria. Regarding the outcome measures, the only framework for quantifying and categorising phonemic errors in dysarthric speakers that has received reasonable attention in the literature is Kent et al.'s (1989) method of phonetic-contrast analysis. Despite the fact that this technique has been implemented in several research studies, it has not been subject to rigorous investigation to assess the validity of its underlying assumptions. Therefore, the main goal of this thesis was to address this gap in the literature. Specifically, the following research questions were identified (Sections 2.1.5

and 2.1.6), with the proviso that the population sample should not include speakers of very low intelligibility:

Q1) Is the range of phonemic-substitution errors typically observed in Belgian Dutch speakers with dysarthria adequately represented by a reasonable number of phonetic-contrast categories?

Q2) Is there close agreement between the phonetic-contrast error profiles identified by (a) different listeners and (b) the same listener on different listening occasions?

Q3) What is the threshold for detecting dysarthria using single-word intelligibility testing?

Q4) Are there phonetic-contrast categories that should be excluded from a clinical assessment of Belgian Dutch dysarthria because they are equally vulnerable in neurotypical speakers?

Q5) Are there significant differences between the word-accuracy values and error profiles yielded by an open and a closed listener response format?

Q6) Are the number and types of error identified by phonetic-contrast analysis predictive of real-world intelligibility?

Q7) What are the phonemic and phonetic-contrast errors of Belgian Dutch speakers with dysarthria?

The above research questions led to the design of this project. The following two paragraphs outline the general methodology. This is followed by four subsections, each of which describes one of the four studies that comprised this thesis. Each subsection begins with a brief summary of the gap in the literature that informed the study design. The specific objectives of the study are then presented, where **bold** font is used to denote technical objectives and research questions of a broad or preliminary nature, while *italics* denote testable hypotheses and research questions with a narrow focus (i.e., meaning that they are framed in such a way that a definitive answer will be obtained).

The project began by **developing a novel phonetic-contrast assessment** that was modelled on Kent et al.'s (1989) test, but is applicable to Belgian Dutch speakers from the Antwerp region. In addition, the proposed assessment was designed to test all phonemes of Dutch on at least three occasions. This meant that it would be capable of providing information about vulnerable phonemes as well as vulnerable phonetic contrasts, which was considered important given the limited prior research on articulatory errors in Dutch dysarthria. A further difference with respect to Kent et al. (1989) is that the distribution of phonemes was chosen to be reasonably representative of that used in everyday speech, to increase the likelihood that the overall word accuracy would be strongly associated

with real-world intelligibility. As documented in Chapter 3, as a result of these requirements, the development of the word list was a significant undertaking and can be considered an important methodological contribution of the thesis in its own right.

The test was then administered to two groups of speakers from the Antwerp region: (i) individuals with acquired dysarthria due to a variety of neurological conditions (n = 10); (ii) age-matched control speakers with no known neurological impairment (n = 8).[18] A series of listening studies was carried out to improve understanding of the nature and value of the information provided by phonetic-contrast analysis. In particular, these studies addressed the seven questions listed above to an extent that was possible given the available time frame and resources. The listeners were native speakers of Belgian Dutch from the Antwerp region, most of whom had no formal training in listening to or transcribing dysarthric speech.

2.5.1. Study 1: Transcription of single words uttered by speakers with dysarthria

Kent et al.'s (1989) test employs a multiple-choice response format, a consequence of which is that the test is only able to identify errors that are represented in the list of distractors. In particular, the minimal-pair distractors were chosen such that they each constitute an error in a single phonetic feature for one of the three word segments. However, as shown in Section 2.1.5, there does not appear to be any direct evidence to support the assumption that the phonemic-substitution errors typically observed in speakers with dysarthria are adequately represented by a reasonable number ($\lesssim 20$) of phonetic-contrast categories. Furthermore, to the best of the author's knowledge, for the current language of interest (Belgian Dutch), there are in fact no studies that report the types of phonemic or phonological errors yielded by speakers with dysarthria as judged by perceptual analysis.

In light of the above, the overarching goal of the first study was to obtain cross-linguistic evidence for the feasibility of analysing dysarthric speech based on phonetic-contrast analysis. More specifically, it was expected that by identifying the full range of speech errors that listeners perceive in unconstrained conditions (orthographic transcription), it would be possible to **assess whether it is justified to confine the perceived errors to a reasonable number of phonetic-contrast categories** (Q1). The second aim of Study 1 was to **obtain preliminary data regarding the phonemic and phonetic-contrast errors observed in Belgian Dutch speakers with dysarthria** (Q7). Even preliminary

[18] Ten participants were recruited, but the data from two of them proved to be unusable; see Chapter 5.

data (i.e., obtained from a small number of speakers and with a newly developed, potentially suboptimal word list) could yield important insights: firstly, from a methodological perspective, the findings could inform the choice of speech materials in future dysarthria assessments, including refinement of the currently proposed test. From a theoretical perceptive, the identification of prominent articulatory errors could improve understanding of the underlying neuromuscular deficits in acquired dysarthria.

The third aim of Study 1 was to **obtain preliminary evidence regarding the inter-rater reliability of phonetic-contrast analysis by means of orthographic transcription**. The evidence in relation to this question is referred to as preliminary because ideally, the reliability metric would have been calculated for a set of trained SLTs (to maximise external validity). Unfortunately, it was not possible to recruit sufficient numbers of listeners who met this criterion, and as mentioned, the majority of listeners had no formal experience in disordered speech. A second reason for considering this evidence to be preliminary is that the level of agreement should have been calculated for the final outcome measure – the *contrast-error profile*. This was not possible, due to limitations of the listening trials (see Chapter 3), meaning that the level of agreement was calculated for responses to *individual words*. These two reliability metrics could be different, as different listeners might perceive the same contrast error for different targets. A third reason for regarding these findings as preliminary is that an inter-rater agreement metric derived from a novel assessment may not reflect the reliability of the technique once it has been optimised and validated. In addition to the fact that the *inter*-rater reliability measure was nonideal, there were insufficient resources to carry out repeat listening occasions in this thesis, meaning that it was not possible to examine *intra*-rater agreement. Therefore, a thorough investigation of Q2 is left for future research.

Study 1 was exploratory in nature, for a number of different reasons. Firstly, there was insufficient prior knowledge to formulate testable hypotheses or measurable research questions. Secondly, the problem is multifaceted in the sense that there may be a number of different barriers to the process of reducing the observed errors to a manageable number of phonetic-contrast categories (and all of these potential barriers, to reiterate the first point, are under-researched). For example, it could be the case that a large proportion of the observed errors involve contrasts in more than one phonetic feature simultaneously (equivalent to a *feet* - *meet* error in English). A further possibility is that the number of distinct contrast categories turns out to be too large to render the technique feasible. As discussed in Section 2.1.5, this could arise for a number of reasons, including (i) the discovery of categories that were not included by Kent et al. (1989) and

(ii) the need to subdivide some of the Kent categories based on evidence that they represent more than one type of underlying articulatory deficit. The third challenge with formulating hypotheses is that it would be difficult to quantify the outcomes in such a way as to draw definitive conclusions about the usefulness of the technique. For example, what would be an 'acceptable' value for the proportion of phonemic-substitution errors that can be described in terms of a contrast in a single phonetic feature? And how many categories would be considered 'too many' for the technique to be infeasible?[19]

2.5.2. Study 2: Transcription of single words uttered by neurotypical speakers

An implicit assumption of Kent et al.'s (1989) approach is that all errors identified in a dysarthric speaker are due to impaired speech production. As shown in Section 2.1.5, the evidence base for this assumption in the specific case of Kent et al.'s word list is limited. From a broader perspective, several studies have provided normative data for the number (and sometimes the types) of phonemic errors observed in single-word production tasks in English. However, these studies have produced mixed findings, with some studies yielding almost no errors and others reporting word accuracies below 80%. For the Dutch language, two studies with relatively large sample sizes reported phoneme accuracies below 85% in some speakers, as yielded by the NSVO (de Bodt et al., 2006). Therefore, it seemed unlikely that perfect or near-perfect accuracy would be observed in control speakers in the present study, leading to the formulation of Q3:

> *What is the threshold for dysarthria detection in Belgian Dutch speakers from the Antwerp region based on metrics of intelligibility derived from single-word reading?*

It was not possible to make predictions about whether any of the phonetic-contrast errors observed in Study 1 would be equally common in neurotypical speakers (Q4). A possible exception to this statement is in relation to the phonological processes mentioned in Section 2.1.5, whereby the Antwerp accent has very similar average formant values for [i] and [ɪ] as well as for [a] and [ɑ]. However, these confusions refer to specific phonemes. Therefore, if they are equally vulnerable in neurotypical speakers, then this would merely imply that these particular vowel-pairs are not suitable test stimuli, *not* that the

[19] This is not a straightforward question, as the number of stimuli that would need to be tested in order to yield a reliable estimate of the error rate for a given contrast category is unknown. Bunton and Weismer (2001) found that of the 13 word pairs used to test the high-low vowel contrast in Kent et al.'s (1989) assessment, four word pairs did not produce an error in any speaker. Therefore, a considerable amount of research is still needed, even for the English version of the test, to understand the interaction between the test stimuli and a speaker's propensity for errors.

categories they represent are irrelevant.[20] Since there was insufficient evidence to formulate any hypotheses, this objective was left as a research question:

Do any of the phonetic-contrast categories identified in Study 1 yield errors rates that are not significantly higher in speakers with dysarthria than in neurotypical speakers?

2.5.3. Study 3: Multiple-choice identification of phonetic-contrast errors in speakers with dysarthria

It was argued in Section 2.1.5 that phonetic-contrast analysis should ideally be conducted using a closed-response format, as this would significantly reduce the burden on the assessor. However, there was concern that this may introduce bias into the listeners' responses compared to an open format where errors on multiple segments may be recorded simultaneously. The main goal of Study 3 was to address Q5: Are there significant differences between the word-accuracy values and error profiles yielded by an open and a closed response format? While a number of studies have examined the first part of this question (regarding word-accuracy scores), there is almost no literature on how the response format affects the speaker's profile of errors.

Due to the limited number of listeners, multiple-choice data were acquired for speakers with dysarthria only. The study had four objectives. The first related to the word-accuracy scores yielded by the two response formats. The existing evidence seemed to suggest that, on average, a closed response format results in higher word accuracy than an open format. Therefore, the first objective was to test the following hypothesis:

Intelligibility metrics derived from single-word reading are higher for the forced-response mode than the free-response mode.

Secondly, it is of interest to examine the *consistency* of the relationship between the two intelligibility metrics across speakers. If the relationship is not consistent, then the implication is that the two response formats are qualitatively different, in which case further research would be needed to determine which format produces intelligibility scores of greater functional relevance. There are a number of outcome measures that could be used to measure consistency, such as the correlation coefficient between the two sets of accuracy scores, the variability in the difference between the two accuracy scores across the cohort, and the order in which the speakers are ranked using each score. Yet no matter which of these outcome measures is used, there are no clear guidelines for

[20] As shown in Chapter 4, it turned out that it was not possible to reduce vowel errors to phonetic-contrast categories of the type proposed by Kent et al. (1989), e.g., high vs. low vowel. Therefore vowel confusions were described in terms of phonemic substitutions. However, this was unknown at the time of the design of the study.

defining a level of consistency that would be considered acceptable in terms of regarding the two methods as interchangeable. Furthermore, as mentioned, the small sample size limits the generalisability of the findings. For these reasons, the second objective was left as a broad question:

What is the consistency of the relationship between word-accuracy scores for the free- and forced-response modes?

The third objective related to Q2: What are the levels of intra- and inter-rater reliability for phonetic-contrast analysis? As was explained for Study 1, a thorough investigation of this question was not possible within the confines of the present thesis: intra-rater reliability could not be assessed at all and the measure of inter-rater reliability was nonideal. Given these limitations, the third objective was expressed as follows:

To obtain preliminary data regarding inter-rater reliability in a forced-choice single-word intelligibility test of Belgian Dutch speakers with dysarthria.

The final objective was to compare phonetic-contrast error profiles for the two response modes. The evidence produced by Bunton et al. (2007), along with the theoretical arguments presented in Section 2.1.6, suggested that some substantial differences would arise. However, the error rates produced by the two techniques are conceptually different; in a free-response study, the denominator (the number of occasions on which each phonetic contrast is tested) is unknown, whereas in a forced-response mode, it is dictated by the list of distractors. For this reason, a direct quantitative comparison of the two methods was not possible, and it could not, for example, be ascertained whether the error rates for a given contrast category were statistically significantly different in the two response modes. Instead, this fourth objective was investigated by means of correlation analysis between the two sets of ranked errors, for which the null hypothesis states that the two sets of observations are independent (i.e., a correlation of zero). The interest was in determining whether, in fact, the correlation between the two sets of rankings *exceeds* zero, leading to the following alternative hypothesis (right-sided):

The degree of correlation between the ranked errors yielded by the two response modes exceeds zero, both in the case of individual speakers, and when error ranks are summed over the cohort.

The alternative hypothesis was directional because it was considered highly implausible that there would be a significant negative correlation between error rankings in the two response modes.

2.5.4. Study 4: Correlation between single-word intelligibility and spontaneous-speech intelligibility in speakers with dysarthria

The final question to emerge from the literature review was whether the number and types of error identified by phonetic-contrast analysis are predictive of real-world intelligibility (Q6). As discussed in Section 2.3, addressing this question requires an explanatory approach and would have been a significant undertaking – one that was beyond the scope of the present thesis. Therefore, it was decided that a preliminary investigation would be carried out, for the speakers with dysarthria only, to examine the degree of correlation between an intelligibility measure derived from single-word reading and an intelligibility measure derived from spontaneous speech – a form of connected speech that has not been widely investigated in previous studies of this nature.

Regarding hypothesis formation, it was noted that the most closely related study from a methodological perspective, that of Lagerberg et al. (2014), reported a Pearson's correlation coefficient of 0.79 ($p < 0.01$) for the relationship between the percentage of correct consonants in single words and a measure of SSI in children. However, it was argued that it would be difficult to use this as a basis for forming a hypothesis about the level of correlation in the current study, as there are likely to be substantial differences between the two populations in terms of factors such as the coherence of their discourse, utterance length, lexical and grammatical complexity, and the extent to which they attempt to use compensatory strategies. Nevertheless, when the full body of evidence in this literature review is considered, including the findings of de Bodt et al. (2002) and Rong et al. (2016), it seems reasonable to hypothesise that a moderate correlation level, of at least 0.5, will be observed. Unfortunately, it was unlikely that the present study (n = 10) would have sufficient power to test this hypothesis. For example, when n = 10, a Pearson's *r* of at least 0.86 would need to be observed to yield a lower-bound confidence interval in excess of 0.5. Therefore, the first objective was expressed as a question:

> *What is the level of correlation between metrics of intelligibility derived from single-word reading and a metric derived from spontaneous speech?*

A secondary objective of Study 4, as discussed in Section 2.4, was **to assess the suitability of the Lagerberg et al. (2014) metric for quantifying SSI in speakers with dysarthria**. This objective was exploratory in nature; thus no specific hypothesis or question was posed. It was expected that even though the study is limited to assessing one particular SSI metric, the findings should have broader relevance in terms of improving understanding of some of the methodological issues surrounding the quantification of intelligibility in spontaneous speech.

3. Methods

This chapter describes the four main aspects of data acquisition: (1) the participants, (2) the interview procedure, (3) the speech tasks and (4) the listening studies. To maximise the future value of the data, the interview included a wide range of speech tasks: single-word reading, picture naming, sentence reading, the delivery of a monologue, and the reading of a short passage. However, only the speech stimuli analysed in the present thesis (single words, pictures[1] and the monologue) are described below. Ethical approval for the study was granted by the Ethics Committees of the School of Health Sciences, City, University of London (UK) and the Middelheim Hospital, Antwerp (Belgium). Participants were provided with a Participant Information Sheet (Appendix 1) and gave notification of their informed consent by signing the Participant Consent Form (Appendix 2).

3.1. Participants with dysarthria

3.1.1. Recruitment

Participants with dysarthria were recruited by clinical staff within the Antwerp hospital network, known as the ZNA (Ziekenhuis Netwerk Antwerpen). The inclusion criteria were as follows:

- Diagnosis of acquired dysarthria due to either (a) a cerebrovascular accident (CVA) or (b) neurological damage or disease that mainly affects the cerebellum
- At least 4 weeks post-injury (acute) or post-diagnosis (chronic)
- Native speaker of Belgian Dutch
- Greater than or equal to 18 years of age (no upper age limit)

The diagnosis of dysarthria was not based on specific assessments or scores, as it was thought that imposing such criteria would have significantly reduced the sample size. This is because the prior experience of the lead clinician indicated that the rate of uptake in research studies was often highest during the early stages of the patient's involvement with the SLT team. Therefore, it was deemed prudent to carry out the interview at the earliest possible stage, sometimes before any formal assessment by a speech and language therapist (SLT) had been carried out (and in some cases, no such assessment was ever performed, due to the perceived greater importance of other clinical activities). Nevertheless, in all cases, the participant had at least undergone an initial consultation with

[1] The picture-naming data were not analysed routinely (i.e., for each speaker) and do not constitute an important part of this thesis. However, some preliminary findings for this task are mentioned in Chapter 8.

a neurolinguist or an experienced SLT. These clinicians were instructed to consider an individual "dysarthric" if their speech was judged to be not fully intelligible to an unfamiliar listener. Thus, if an individual had disordered speech characteristics, but remained, in the clinician's estimation, 100% intelligible in conversational speech, then they were not recruited for the study.

As can be seen, the study recruited participants from among two aetiological groups: (1) non-cerebellar CVA and (2) cerebellar injury or disease (including cerebellar stroke). The motivation had been to compare profiles of phonetic-contrast errors for the two clinical populations. Unfortunately, due to unforeseen circumstances, it was not possible to recruit participants in the manner originally planned, resulting in a lower recruitment rate than that required to make a between-group comparison. As stated in Chapter 2 (Section 2.5), the vast majority of questions addressed in this research project were methodological and not contingent on the presence of two clinical groups. The aim of conducting a between-group comparison was modified to the following: to obtain preliminary data regarding the phonemic and phonetic-contrast errors perceived in Belgian Dutch speakers with dysarthria.

Participants were required to be four weeks post-injury or diagnosis, as this increased the likelihood that they would be in a reasonably stable neurological, physical and psychological state. This criterion was mainly included for ethical reasons (to reduce participant vulnerability), rather than to achieve a more homogeneous sample. Indeed, due to the low anticipated participant numbers, it was decided that it would be necessary to include participants who were inpatients at one of the Antwerp hospitals, meaning that it was likely that some of these individuals would be exhibiting signs of stress, fatigue and perhaps even confusion. For this reason, the clinicians responsible for recruiting participants were instructed not to approach individuals who did not appear to be in a sufficiently stable physical and psychological state to conduct the interview. Furthermore, while the interview was being conducted, the author regularly asked participants whether they were coping and reminded them that they may stop or pause at any time.

The study only recruited speakers whose native language was Belgian Dutch. This is because different speech characteristics might be observed in people for whom Dutch is a second language. Similarly, variation in accent could be a confounding factor. However, rather than imposing accent requirements during recruitment, it was decided that all accents of Belgian Dutch would be permitted, and information about accent would be gathered as part of the case history. Depending on the number of participants and the observed variation in accent, it would then be decided whether to subdivide the analysis

on this basis. However, it transpired that all participants had lived in the Antwerp region for the majority of their adult lives and were identified by an expert (Jo Verhoeven) as having an Antwerp accent.

The exclusion criteria for participation in the study were as follows:

- Pre-injury history of a significant developmental, motor or neurological disorder (e.g., epilepsy, autism, cerebral palsy)
- Pre-injury history of a significant speech disorder (due to, for example, hearing impairment or cleft palate) or a significant language disorder (e.g., aphasia, AOS)
- Moderate / severe acquired aphasia or apraxia of speech
- Moderate / severe cognitive impairment or mental health impairment

Due to ethical rules that prohibit non-clinical researchers from consulting a patient's medical records before they have consented to participate in the study, the clinicians responsible for recruitment were requested to ensure that, to the best of their knowledge, potential participants met the recruitment criteria. This judgment was based on personal familiarity with the patient and their case history. The latter included, as standard, the results of the Mini Mental State Examination (MMSE), where the minimum score for inclusion in the study was 24/30. The exclusion criteria were chosen on the basis that the condition in question would either (a) render the process of data acquisition difficult or impossible (e.g., moderate / severe aphasia), or (b) introduce an unreasonable amount of additional noise into the data (e.g., conditions such as cerebral palsy result in their own particular speech errors). The level of cognitive, language or mental-health impairment was considered to be "mild" if it did not, as far as the clinicians were able to judge, impair the participant's ability to carry out the required speech tasks. The decision to allow mild cognitive impairment and/or aphasia was based on the clinical reality that pure dysarthria (i.e., in the absence of any other neurological impairment) is relatively uncommon, especially in stroke (Duffy, 2005: p.258).

3.1.2. Sample size

The plan was to recruit 20 neurological participants, 10 with dysarthria due to non-cerebellar stroke and 10 with dysarthria due to a cerebellar condition. This sample size was chosen to be similar to or larger than that used in previous studies in which phonetic-contrast analysis had been employed to yield prototypical error profiles for specific clinical populations (e.g., Blaney & Hewlett, 2007; Bunton & Weismer, 2001; Kent et al., 1990; Haley et al., 2000; Whitehill & Ciocca, 2000b). Within each aetiological group, the aim was to include approximately equal numbers of male and female subjects, as previous studies have suggested that speech-error profiles may be sex-specific (e.g., Bunton & Weismer, 2001;

Riddell et al., 1995; Kent et al., 1992). However, as explained above, the planned recruitment method could not be implemented. Once this became apparent, the goal was to recruit as many participants as possible, irrespective of their sex and aetiology (either CVA or cerebellar disease),[2] within the time available. The number of participants with dysarthria was 10. The following subsection describes their clinical characteristics.

3.1.3. Sample characteristics

The personal and medical data gathered for each participant are presented in Table 3.1. All variables were chosen on the basis that they might be expected to influence an individual's single-word or spontaneous-speech characteristics and thus might need to be taken into account when analysing the data. In accordance with ethical rules, the author was only authorised to consult the hospital records after the participant had given informed consent, which took place at the beginning of the interview. In most cases, the participant did not display any obvious difficulties, and subsequent consultation of the hospital notes confirmed that no relevant difficulties were present. However, for three individuals, it became apparent during the interview that they had co-occurring conditions that either (a) limited their ability to carry out some of the speech tasks or (b) demonstrated a more widespread region of damage than that suggested by neuroimaging. The specificities of these cases are described later in this subsection.

The clinical and personal data pertaining to each speaker are shown in Table 3.2. Herein, a specific speaker is referred to by the abbreviation "S" followed by the identification number indicated in the table. Table 3.2 shows only the most important characteristics from Table 3.1. However, a column has been included (Column 6) to record characteristics that were of particular relevance for a given speaker. The final column shows word accuracy in the single-word reading task as assessed by orthographic transcription. All but one of the speakers were strongly right-handed. The remaining speaker (S4) performed some tasks with her left hand, but wrote with her right hand. Many of the participants with longstanding dysarthria had received speech therapy aimed at improving articulation, including the pronunciation of specific phonemes (in particular, /l/, /r/ and consonant clusters). There was also emphasis on improving spontaneous speech, with "reduced speech rate" being the most common strategy. Therapy in the chronic stage consisted of weekly sessions that sometimes continued for a period of several years.

[2] The possibility was considered of expanding the recruitment criteria to allow for the inclusion of participants with aetiologies other than CVA and cerebellar disease, but this would have required an amendment to ethics approvals at two different institutions, which would have been too time-consuming given the remaining time available on the project.

Patient characteristic	*Source(s) of information*
Age	Hospital records
Sex	Interview
Place of upbringing (town / city)	Interview (written question)
Place of residency during past ten years	Interview (written question)
Mother tongue	Interview (written question)
Languages spoken other than Dutch	Interview (written question)
Hand dominance (Edinburgh Handedness Inventory; Oldfield, 1971)	Interview (self-reported answers)
Medical diagnosis, including results of computed tomography (CT) and magnetic resonance imaging (MRI)	Hospital records
Time since injury / diagnosis	Hospital records
Medication that can affect speech	Interview (written question)
SLT treatment already administered for dysarthria and any other relevant information from SLT sessions	Hospital records
Any other relevant information, especially cognitive / visual difficulties	Observation during the interview and/or hospital records

Table 3.1. List of participant characteristics recorded in this study.

ID	*Diagnosis*	*M/F*	*Age*	*Disease duration*	*Other relevant information*	*% correct words*
1	Amyotrophic lateral sclerosis†	F	70	Several weeks		78.0
2	Ischemic CVA (suspected). No clear lesion on imaging; microvascular damage and iron deposition in brainstem	F	87	4 weeks (sudden onset)	• Mild right-sided paresis (loss of normal arm swing and slight circumduction)	77.3
3	Surgical damage from medulloblastoma, left cerebellum. Recent imaging showed no new damage at time of interview, but there was later evidence of	M	35	62 mths	• Diffuse cognitive deficits (verbal IQ, concentration, working memory, word-finding) • Epileptic fits, intention tremor	73.2

	tumour recurrence and metastasis.				• Smooth gait, but unstable in straight-line test; other ataxic signs, e.g., knee-heel and Romberg tests	
4	Surgical damage from hemangioblastoma in fourth ventricle and craniocervical junction. T2 hyperintensity on MRI in posterior cerebellum.	F	56	22 mths	• Right-sided arm paresis and tongue paralysis • Gait "wooden", unstable and slow; unable to walk in straight line	49.1
5	Progressive genetic condition causing cerebellar atrophy (visible on MRI)	M	63	> 5 years (gradual onset)	• Ataxic gait	70.6
6	Ischemic CVA (right cerebellum)	M	45	4 mths	• Right-sided paresis • Ataxic gait and coordination disturbances	83.9
7	Ischemic CVA (pons / left cerebral peduncle)	F	75	9 mths	• Right-sided paresis • Reduced balance	88.4
8	Ischemic CVA (right-sided, cortical "watershed" stroke in border zone of posterior and middle cerebral arteries)	M	81	4 weeks	• Mild cognitive deficit, lability, left-sided neglect • Left-sided facial paresis & weakness • History of vocal fold hyperkeratosis	68.8
9	Ischemic CVA (left cerebellum)	M	63	78 mths	• Bilateral pseudoexfoliation syndrome and scotomas • Ataxic gait; balance and coordination difficulties	62.0
10	Amyotrophic lateral sclerosis† (suspected)	M	68	Several weeks		90.7

† These speakers were suspected of cerebellar disease at the time of the interview. S1 was diagnosed with ALS over a year later. The diagnosis for S10 remains unconfirmed at the time of writing. Since there was no sudden onset for these participants, the precise duration is not stated.

Table 3.2. Personal and clinical information pertaining to the participants with dysarthria.

Three of the participants had difficulties that influenced their ability to carry out the speech tasks. Participant 3 was a 35-year old gentleman who had had two previous occurrences of a medulloblastoma that had left him with a light dysarthria. He had recently been readmitted due to new symptoms, most notably, epileptic seizures. At the time of the interview, a recent MRI showed sequelae of the surgery carried out on the earlier tumours (in the left cerebellum), but no conclusive evidence of a recurrence or of any other neuropathology (e.g., encephalitis or meningitis). During the interview, it became apparent that the participant had cognitive impairment of a more widespread nature than would normally be expected due to a purely cerebellar condition – in particular, executive dysfunction.[3] For example, the picture-naming task had to be abandoned because he was unable to follow the instructions that all pictures represented words of one syllable only and that the words must be uttered without any prefixes (e.g., the definite or indefinite article). Following the interview, consultation of the participant's hospital records revealed that he had diffuse cognitive deficits, with abnormal scores in most subtests of the Repeatable Battery for the Assessment of Neuropsychological Status (R-BANS) and the Wechsler Adult Intelligence Scale IV (WAIS-IV). His performance in both the picture-naming task (of the R-BANS) and the word-definition task (of the WAIS-IV) was abnormal. The hospital notes also revealed that he was experiencing word-finding difficulties and was scheduled for an aphasia assessment, although no record of such an assessment could be found. Unfortunately, the participant's health deteriorated rapidly in the weeks following the interview, due to a confirmed recurrence of his brain tumour, which might explain why further SLT assessment was not performed. Regarding the participant's eligibility for the study, he was able to carry out the speech tasks relevant to the current thesis (single-word reading and delivering a monologue). His monologue, which was on the subject of his illness, was reasonably fluent and coherent, although there were occasional pauses in unusual places, which may have been due to word-finding difficulties. His clinical presentation showed clear signs of ataxia, including bilateral ataxic knee-heel tests, instability when trying to walk in a straight line, and a positive Romberg test. However, given his other clinical features, combined with the fact that it later transpired that his tumour had recurred and metastasised, it is possible that his site of neurological injury had extended beyond the cerebellum at the time of interview.

Participant 8 was an 81-year old gentleman who had been admitted to hospital four weeks prior to the interview following symptoms indicative of a stroke, including left-sided arm

[3] Note, however, that Schmahmann and Sherman (1998) described a cluster of multimodal cognitive disturbances in patients with focal cerebellar lesions, a phenomenon that they refer to as "cerebellar cognitive affective syndrome".

and leg weakness, left-sided facial paralysis, and slurred speech. His MMSE score of 24/30 suggested mild cognitive impairment. A CT-scan taken two months after the interview was reported to be consistent with "a recent non-haemorrhagic CVA in the border zone vascularised by the middle cerebral artery and the right-sided posterior cerebral artery". During the interview, several difficulties came to light that affected the participant's performance on the speech tasks. Firstly, during the monologue (where he talked about his recent hospitalisation), he showed emotional lability that translated into fluctuating intelligibility. More specifically, when he became emotional, his speech took on different perceptual characteristics (in particular, a much higher pitch) and became highly unintelligible. Secondly, during the sentence-reading task, which was administered using PowerPoint, it became apparent that the participant had a left-sided neglect, as he only uttered the second half of each sentence. Furthermore, the speaker's behaviour and interaction with the author showed signs of executive dysfunction. Following the interview, consultation of the hospital notes confirmed these observations. The report from the occupational therapist described the patient as "emotionally uninhibited" with "reduced problem-solving abilities". Although he was observed to be independent with regards to self-care (washing and dressing), he was also "chaotic", "[did] not always follow a logical order" and failed to notice things in his environment. The SLT reported that he was "sometimes very confused" and "clearly [had] no insight into his illness". A marked left-sided visual neglect was also mentioned. Regarding the suitability of the participant for the present study, his left-sided neglect did not appear to affect his single-word reading. His monologue seemed to be reasonably logical and fluent (although perhaps less so than most of the other participants). However, as mentioned, there were breakdowns in intelligibility when he became emotional and he did not seem to show insight into his communication difficulties, nor to check for signs of comprehension on the part of the author. Thus, there was no attempt at repair, nor any obvious indication of the use of compensatory strategies.

Finally, Participant 9 (male, aged 63) had experienced a cerebellar CVA almost 7 years prior to the interview, which had left him with a permanent, moderate dysarthria and classical signs of cerebellar damage, such as dysdiadochokinesia, scanning speech and an ataxic gait. The author was not made aware of any other difficulties prior to the interview. However, during the interview, it became apparent that he had some visual difficulties that caused reading problems. These were not consistent; the majority of words appeared to be read correctly. However, approximately 20% of words had to be discarded because they were misread rather than (or in addition to) being mispronounced. The misread words were clearly identifiable because they did not correspond to the typical speech errors perceived in this speaker (who had a relatively consistent dysarthria). Rather, there was close

similarity between the *orthographic* appearance of the uttered word and that of the target. In addition, the uttered word was generally of higher frequency than the target. A typical example was the target *haard* (/ha:rt/ - 'hearth') produced as the higher-frequency word *baard* (/ba:rt/ - 'beard'). Occasionally, the speaker realised his error and self-corrected. The correct realisations were included in the study provided they did not appear to have been pronounced in an unnatural manner (e.g., over-articulated). If there was even a small suspicion that the word had been misread, it was discarded. Following the interview, consultation of the patient's notes revealed that a few weeks previously, he had undergone a thorough examination at an eye clinic. He had previously been diagnosed as having bilateral pseudoexfoliative syndrome, a condition characterised by the deposition of a protein-like material within the eye, which increases the risk of developing glaucoma. At the appointment, he had complained of having experienced epiphora (watering) of the left eye for a period of a week. The examination revealed bilateral visual-field scotomas (drop-outs) that were not suggestive of glaucoma, but perhaps an occipital lobe CVA. The recommendation was for a further CT scan of the brain. However, the patient refused out of fear of the procedure. In the light of this information, one must consider the possibility that this participant's region of damage was no longer confined to the cerebellum (although an additional lesion solely in the occipital lobe would not be expected to introduce any new dysarthric symptoms). Furthermore, as mentioned, the speaker's visual difficulties caused him to misread some of the single-word targets. Despite the author's best efforts to identify and remove these instances, this may not have been achieved with perfect accuracy.

Table 3.2 demonstrates that with one exception (S8), the participants whose dysarthria was due to stroke showed damage in subcortical structures such as the cerebellum and the brainstem. Duffy (2005: p.171) reported that when speakers were categorised according to their dysarthria type, only half of those with a diagnosis of ataxic dysarthria had an identifiable lesion or region of atrophy in the cerebellum; most of the remaining speakers showed lesions of the brainstem or midbrain. Consequently, given that there was only one speaker with CVA whose imaging data did not shown signs of subcortical damage, it was not possible to define a CVA group that could be definitively categorised as 'non-cerebellar' (or 'non-ataxic'). At the same time, there was insufficient justification for considering all speakers with subcortical damage as forming one group. Regarding the remaining two speakers (S1 and S10), one of them had a definitive diagnosis of ALS, while the other had suspected ALS, but this was unconfirmed at the time of writing.[4] Both of these participants

[4] The hospital in which he was interviewed held no further records on him, so it was believed he was being followed up elsewhere.

had recent-onset dysarthria as their only recorded symptom, which, in the case of ALS, is generally a sign of lower motor neuron involvement and hence flaccid dysarthria (Tomik & Guiloff, 2010). In summary, due to the variety of medical diagnoses and sites of damage on imaging, it was decided that the participants would be analysed as one group with dysarthria due to various aetiologies. Nevertheless, it can be seen from Table 3.2 that for four of the speakers (S3, S5, S6 and S9), the underlying condition was mainly or purely cerebellar, at least according to neuroimaging. For this reason, some of the findings in this study, specifically those relating to the speech characteristics of the aforementioned participants, are discussed in the context of existing knowledge about articulatory errors in ataxic dysarthria and the role of the cerebellum in speech.

In addition to their aetiologies and lesion sites, the participants formed a heterogeneous group in terms of factors such as age, gender, presence of co-occurring cognitive deficits, dysarthria severity, and the amount of SLT treatment. Due to the small sample size, it was not possible to examine the individual effect of any of these variables on speech error profiles. To some extent, this inherent heterogeneity is a clinical reality, and a high incidence of co-occurring cognitive deficits has been reported in many previous studies (e.g., Urban et al., 2006).[5] Nevertheless, as described above, some of the deficits observed among the current set of speakers resulted in tangible limitations.

3.2. Neurotypical control subjects

Age-matched, healthy control subjects were included so as to (a) determine a cut-off for the diagnosis of dysarthria by single-word reading accuracy and (b) distinguish between true dysarthric errors and errors that arise due to natural phonological processes or close perceptual similarity. The single-word intelligibility test developed in the present study (i.e., a set of highly confusable, monosyllabic, real words) is the first of its kind in the Dutch language. Therefore, the types and frequencies of phonetic confusions that would be observed in neurotypical speakers were unknown. The study recruited 10 participants who, to the best of their knowledge, had no congenital or acquired condition that would affect their cognition, speech, language or reading ability. They were invited to participate via a site-wide email sent to the University of Antwerp calling for individuals who considered themselves to have an Antwerp accent. As explained in Chapter 5, the data from two of the subjects had to be discarded. The mean age (± 1 SD) of the remaining eight

[5] Note that this statement mainly applies to stroke, as does the reference (Urban et al., 2006). Cognitive deficits in individuals with pure cerebellar disease are thought to be much less common. When they do arise, they have generally been attributed to co-occurring non-cerebellar damage. However, see Footnote 3 and Duffy (2005: p.166) for an alternative viewpoint.

control subjects (4 M, 4 F) was 70.4 ± 9.4, with a range of 56-83 years. Although two of the participants with dysarthria in this study were considerably younger than this age range (see Table 3.2), the goal of acquiring normative data for the *typical* target population was considered to be more important than that of matching the current sample.

3.3. Interview procedure

This was an observational study in which data were collected by means of participant interviews. The purpose of the interview was to obtain a broad spectrum of speech samples, ranging from single-word monosyllabic productions (which are highly controlled and contain no contextual clues) to spontaneous speech (which, although uncontrolled, represents the most natural form of oral communication). The speech tasks were designed such that, based on previous studies with a similar protocol (e.g., Whitehill & Ciocca, 2000a), the expected duration of the interview would be approximately 45 minutes. According to the leading clinician, this was the maximum duration that the participants would be able to comfortably endure, although precautions were taken to ensure that the interview would be stopped earlier if this were in the participant's interest. In fact, the typical duration of the speech tasks was 20-25 minutes, with a further 10-15 minutes required for taking informed consent and gathering background information (see Table 3.1). The length of time occupied by the single-word reading component, which consisted of 125 words (including practice words), was approximately 8 minutes. Despite the monotonous nature of this task, it was well tolerated by the participants and when questioned, they reported that they had not found it arduous. One participant (S2) reported being tired during the interview. This occurred at the end of the final task (the monologue) when the interview was about to draw to a close.

Interviews took place either on hospital premises or in the participant's home, depending on patient status (i.e., in vs. outpatient) and the participant's wishes. When no preference was expressed, the interview was conducted in the participant's home, as the level of background noise tended to be lower in that environment. Although a sealed room was available for use at both hospital sites, it was found that certain types of noise were particularly prone to transmission through doors, walls, ceilings and floors, such as the movement of furniture and the clacking of heels. Furthermore, there were few soft materials in the hospital environment, so background noise tended to be reflected and amplified rather than absorbed. Large blankets were placed over some of the hard surfaces in the interview room in an attempt to counter this phenomenon. Furthermore, the audio recorder was placed on a sound-absorbing mat. Occasionally, a participant was asked to wait or to repeat a word when the background noise level was considered too high.

The equipment used in the interviews included a laptop computer, digital audio recorder, microphone and video camera. The computer was used to display the stimuli in the single-word reading, picture-naming and sentence-reading tasks. Audio data were recorded on the Marantz PMD-660, a portable, solid-state, compact flash audio field recorder. Preliminary investigations showed that data of superior quality, as judged from audio signals and spectrograms in Praat (Boersma & Weenink, 2018), were obtained by using an external microphone rather than the built-in microphone of the Marantz recorder. Following a review of the available microphones within the project's budget, the model chosen was Audio Technica's AT831b wireless clip-on microphone. This was attached to the participant's clothing at the recommended distance (about 6" below the chin) and was coupled to the Marantz recorder using a Klotz M1FM1N0100 Neutrik XLR 3p 1-metre microphone cable. The microphone was operated in the "flat" frequency response mode (no filtering) and was powered by battery.

The first recording (S1) revealed the presence of a constant source of noise in the speech data. Further experimentation showed that this was due to electrical interference from the power cable of the audio recorder. Therefore, subsequent interviews were carried out using the recorder's lithium battery supply as the source of power.[6] This approach had the added advantage that fewer power sockets were needed in the set-up, which proved to be extremely beneficial when interviewing participants in their homes. The final piece of equipment used in the interviews was a portable USB digital video camera (Flip Mino F360 Cisco). This was run off of its USB-chargeable battery and was immobilised using a flexible tripod (GorillaPod). The video data were not analysed in the present study, due to both limited time and a limited supply of observers. However, the data are available for future analysis. Previous studies have shown that transcription accuracy is greater for a dual (auditory + visual) mode of presentation than for auditory only (e.g., Keintz et al., 2007; Hustad et al., 2007). The combined mode is also more representative of most real-life communicative situations.

The settings of the Marantz recorder were chosen so as to maximise sound quality: ".wav" (i.e., uncompressed) 16-bit digital file format, 48 kHz sampling rate, and a recording level (gain) that was optimised according to the instructions in the Marantz PMD-660 manual. Optimisation of the gain was carried out separately for each participant and speech task. The recording level for single-word reading was optimised based on a set of practice words chosen such that they contained loud phonemes such as sibilant fricatives and vowels. This

[6] Tests also revealed that the signal-to-noise ratio degraded with battery age. Therefore new batteries were used on each recording occasion.

reduced the likelihood that the subsequent test stimuli would contain phonemes of higher amplitude, which would then be at risk of saturation. For the monologue, the recording level was set during the first few seconds of speech, as some speakers found this task to be quite challenging. Therefore, it would have been unreasonable to expect them to deliver a practice monologue solely for the purposes of optimising the gain.

3.4. Speech data

Participants were given instructions both in writing and verbally. Instructions specific to individual speech tasks are described below. For all tasks, the participants were asked to read or speak in a natural manner. Occasionally, a participant was asked to repeat a stimulus on the grounds that the recording conditions were significantly below optimum, for example, due to an inappropriate gain setting or an abnormally loud environmental noise. The interview consisted of the following tasks, listed in the same order as they were presented to the speakers: single-word reading, picture naming, the reading of semantically implausible sentences, the reading of a short passage, and the delivery of a monologue. Before describing the speech stimuli investigated in this thesis (single-word reading and a monologue), a brief overview of the Dutch phonological system is provided.

3.4.1. Belgian Dutch phonology

Figure 3.1 shows the consonants of Belgian Dutch. The voiced obstruents occur only in initial and medial positions; word-final obstruents are always realised without voicing, even when orthography uses the grapheme commonly associated with the voiced form (e.g., *bed*, meaning 'bed', is realised as /bɛt/). Belgian Dutch has both a voiced and voiceless velar fricative. The voiceless counterpart *can* appear in word-initial position, but this is rare and was not tested in the present study. In producing the rhotic, there is free variation between the alveolar and the uvular trill. In contrast to the Netherlands, in Belgian Dutch, the alveolar trill is more common (Verhoeven, 2005). The vowel system of Belgian Dutch is depicted in Fig. 3.2. Note that this representation applies to standard Belgian Dutch, whereas the precise positions and durations of the vowels in any given regional accent may be quite different. For example, Antwerp [ɪ] is very close to [i], and [ʏ] is close to [y]. This issue is revisited in Chapters 4-6 when analysing the vowel confusions perceived in both speakers with dysarthria and neurotypical subjects.

	Bilabial	Labio-dental	Alveolar	Post-alveolar	Palatal	Velar	Uvular	Glottal
Plosive	p b		t d		(c)	k (g)		(ʔ)
Nasal	m	(ɱ)	n		(ɲ)	ŋ		
Trill			(r)				ʀ	
Fricative		f v	s z	(ʃ) (ʒ)		x ɣ		ɦ
Approximant	w				j			
Lateral approximant			l		(ʎ)			

Figure 3.1. The consonants of Belgian Dutch (reproduced with permission from Verhoeven, 2005). The sounds in parentheses either result from surface phenomena or occur only in loan words.

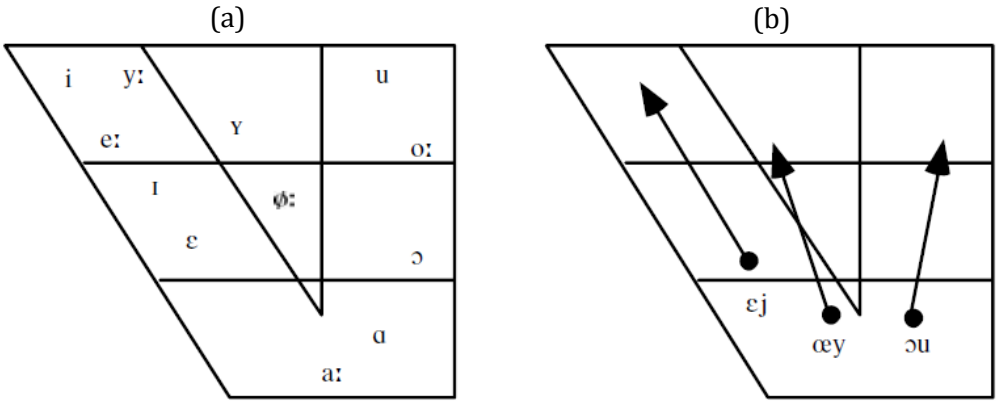


Figure 3.2. The vowels of Standard Belgian Dutch, showing (a) monophthongs and (b) diphthongs (reproduced with permission from Verhoeven, 2005).

3.4.2. Single-word stimuli

This section begins by explaining the methods used to develop the word list, which was the first objective of this study (see Chapter 2, Section 2.5). It then describes the two speech tasks that made use of the single-word stimuli: word reading and picture naming.

The first decision that needed to be made was the total number of words. Prior studies involving speakers with dysarthria have typically tested 70-100 tokens in the single-word production task (e.g., Blaney & Hewlett, 2007; Whitehill & Ciocca, 2000a). Whitehill and Ciocca (2000b) reported that their subjects with cerebral palsy were able to read single words in an intelligibility test at a rate of approximately one word every 5s, in which case 120 words would be completed within 10 minutes. This task duration seemed to be appropriate for the present study, so an upper limit of approximately 120 words was chosen. In reality, the single-word reading task was usually completed in 8-9 minutes.

As stated in Section 2.5, it was decided that the phonemic composition of the word list would match that used in everyday language. However, rather than applying this strategy

to the entire list, a small proportion of the word list (approximately 15%) was reserved to test some of the rarer phonemes of Dutch more frequently. The rationale was that sacrificing a relatively small number of tokens from the phonemically-balanced word list (referred to herein as the "core" list) may result in substantial benefit in terms of gaining information about the speakers' deficits. For example, the voiced glottal fricative was identified as a vulnerable phoneme in dysarthria by Kent et al. (1989) and the associated error category ('initial /h/deletion') has received high error rates in most studies that used Kent et al.'s methodology (e.g., Blaney & Hewlett, 2007; Bunton & Weismer, 2001; Whitehill & Ciocca, 2000b). Therefore, by including extra tokens of word-initial /h/ (beyond the frequency with which the phoneme occurs in natural language), it would be possible to measure the error rate for this contrast category with reasonable reliability.

Various criteria were taken into account when constructing the word list. Firstly, it was decided to use only monosyllabic words, to maximise the opportunity for errors to be perceived (as polysyllabic words are more distinctive). Although monosyllabic words do not represent the full spectrum of word shapes used in everyday communication, their ecological relevance is perhaps greater than one might imagine. In a database of the 500 most common spoken words of Dutch (taken from "Dutch 101", 2014), it was found that 56.8% of these words contained only one syllable. Furthermore, for many of the polysyllabic words, the meaning of the word was confined to just one of the syllables, while the remaining syllable(s) acted as inflectional or derivational affixes. For example, verb infinitives end in the inflectional suffix *-en* (e.g., *hebben* – to have), which, in many accents of Dutch, is reduced to a schwa.

The development of a word list that is phonemically balanced is not straightforward. There are a number of ways of calculating the frequency with which different phonemes of a language occur. One method would be to transcribe an ecologically-valid speech sample and count the number of times that each phoneme appears. The Corpus Gesproken Nederlands (CGN, 2018), which contains a diverse range of speech samples from both Dutch and Flemish speakers, has been analysed in this way (e.g., Zuidema, 2009; Luyckx et al., 2007). However, the resulting distributions, unsurprisingly, show very high frequencies for phonemes that occur in common function words (such as the definite article) or common function morphemes (such as plural affixes). Given that such phonetic units do not convey much meaning and can often be predicted from contextual cues, it is unlikely that they would cause a substantial reduction in intelligibility if produced erroneously. Furthermore, Zuidema (2009) showed that the 5 most common consonant phonemes of Dutch (/t, n, d, r, s/) accounted for more than 50% of transcribed consonants. Therefore, a

speech test based on a *fully* realistic phoneme distribution would only be able to test a small subset of the consonants of the language with reasonable statistical robustness. An alternative approach would be to examine the frequency with which the phonemes of Dutch appear in a set of commonly used words. In such calculations, the relative frequencies of the words themselves are not considered when calculating phoneme frequencies. Therefore, the distribution of phonemes would less closely approximate that used in everyday language. Nevertheless, this approach has the advantage that phonemes that appear frequently in common function words do not predominate.

In practice, a compromise between the above two approaches was reached. Firstly, the author took into account a number of different phoneme-frequency lists that had been published for the Dutch language based on spontaneous-speech samples (Zuidema, 2009; Luyckx et al., 2007; van Severen et al., 2013; Jongstra, 2003). Secondly, phoneme frequencies were calculated by the author based on two different internet sites that listed the 1000 most common words of Dutch ("Memrise", 2014 and "Dutch 101", 2014). All of the words in these internet lists were included in the phoneme-frequency analysis, regardless of the number of syllables that they contained. For multi-syllabic words, the medial consonants were ignored (as these were not tested in the present study), but all vowels were coded. Consonants were analysed separately according to their position (i.e., word-initial vs. word-final). The phonetic composition of consonant clusters was coded precisely. The word lists only contained the common consonant phonemes of Dutch; i.e., borrowed sounds, such as /ʃ/ and /ʒ/, did not appear within the top 1000 words.

Table 3.3 (Column 3) shows the relative frequencies with which it was decided that the vowels of Dutch would be represented in the core word list, based on the two methods described in the previous paragraph. For vowels, there was reasonably close agreement between the *published* distributions of phoneme frequency (Zuidema, 2009; Luyckx et al., 2007)[7] and the *calculated* distribution based on the most common Dutch words. Therefore, an average of these distributions was used. The frequency of each vowel was converted into a number, based on a total allowance of approximately 100 vowel tokens (i.e., the number of words in the core list). For most of the vowel tokens, a range is shown, either due to the fact that the number was not an integer, or because there was some discrepancy between

[7] These two published lists were in close agreement with each other (to within ~5%) for almost all phoneme frequencies (for both vowels and consonants). The largest discrepancy was for the frequency of occurrence of /j/ (21% vs. 33%). This phoneme occurs at the beginning of the informal form of the second-person pronouns (e.g., /jə/- 'you' singular). Furthermore, in some dialects of Belgian Dutch, the spoken form of /jə/ begins with a uvular fricative. Therefore, depending on the sample of the CGN that was analysed (the person(s) being addressed and the accent of the speaker), one might expect different frequencies of /j/ to be obtained.

the various sources. This allowed for some flexibility when choosing the words, which proved useful given the multiple constraints on the word list. The frequency of schwa was not recorded because it is not relevant to single-word targets. Furthermore, only the three "essential" (Collins & Mees, 2003) diphthongs of Dutch were included. There are other vowels that are sometimes referred to as diphthongs (e.g., /iu/ as in the word /niu/ – 'new'), but they are more accurately described as vowel sequences (Collins & Mees, 2003) and are not particularly common; therefore, they were not included. Furthermore, borrowed vowels, such as /ɔ:/ (e.g., /rɔ:zə/ – 'pink', a word of French origin), were omitted.

Vowel	*Example word and translation*	*# tokens*
/ɑ/	/bɑt/ *bad* 'bath'	13-14
/a:/	/ba:t/ *baat* 'profit'	9-10
/ɛ/	/bɛt/ *bed* 'bed'	9-10
/e:/	/be:t/ *beet* '(I) bit'	9-10
/ɪ/	/bɪt/ *bid* '(I) pray'	7-8
/ɔ/	/bɔt/ *bot* 'bone'	7-8
/o:/	/bo:t/ *boot* 'boat'	7-8
/i/	/bit/ *biet* 'beetroot'	5-6
/u/	/but/ *boet* '(I) atone'	3-4
/ʏ/	/bʏs/ *bus* 'bus'	2
/y:/	/by:r/ *buur* 'neighbour'	1-2
/ø:/	/bø:k/ *beuk* 'beech'	1
/ɔu/	/bɔut/ *bouwt* '(He) builds'	2
/œy/	/bœyt/ *buit* 'booty'	1-2
/ɛi/	/bɛit/ *bijt* '(I) bite'	5-6

Table 3.3. Distribution of vowels for the core word list. The third column shows the approximate number of occasions on which it was decided that each vowel would appear.

For word-initial consonants, a third source of information for estimating phoneme frequencies was also used, namely a standard Dutch-English dictionary (van Dale Uitgevers, 2009). The number of pages occupied by each consonant was counted so as to determine the frequencies of consonants in word-initial position. This approach might be expected to produce different results from the previous two approaches, as it includes words of all frequency levels (not just the most common). It is worth noting that this method of analysis was only possible because there is a reasonably consistent mapping of

graphemes to phonemes for Dutch consonants in word-initial position. There are a few exceptions (for example the grapheme "c" in word-initial position can correspond to the phonemes /k/ or /s/ when it appears in isolation and /x/, /ʃ/, or /tʃ/ when it is followed by the grapheme "h"). However, most of the nonstandard pronunciations correspond to words borrowed from other languages and are relatively infrequent. Therefore, the approach was considered to be sufficiently accurate for the present purposes.

For consonants, it was not straightforward to interpret and integrate the information from the different sources. This was partly because the published phoneme-frequency lists did not always separate the results by consonant position (word-initial vs. word-final). However, it was also a natural consequence of the fact that the three different methods of calculating phoneme frequency provide qualitatively different types of information. Nevertheless, it was possible to make reasoned decisions about the distribution of consonant frequencies to be used in the present study. The decision-making process did not follow a consistent procedure that could easily be coded by an algorithm. Therefore, rather than attempting to describe all the factors that were taken into account, the following paragraphs provide two specific examples to give the reader an impression of how the decisions were reached. In these examples, the rating values are numbers out of 10, which indicate how frequently the phoneme occurred, relative to the most common phoneme at the same word position, *for the same data source*.

The first example refers to the voiceless alveolar plosive. The various sources all agreed that in word-initial (C1) position, /t/ only has a moderate frequency (rating of ~ 4-5 relative to the most common initial phonemes). This frequency rating includes clusters (e.g., /tr/ and /tw/), which together comprise about 15-20% of the words beginning with /t/. Given that all the sources agreed, the frequencies with which word-initial tokens of /t/ and /t/-clusters would be tested (see Table 3.4) were chosen to match these findings. The situation for the word-final (C2) position, where /t/ is extremely common (a frequency rating of 10), was less straightforward. One of the reasons for the high frequency of word-final /t/ is that final consonants are always devoiced. However, it was also noticed that in ~65% of the words ending in /t/ in the lists of common Dutch words, /t/ was part of a consonant cluster and, in many of these cases, it acted as an inflectional morpheme. For example, the suffix *–t* is used to denote the second and third person singular of almost every Dutch verb. These instances of /t/ are not likely to have a substantial impact on intelligibility. Thus, it was decided that the frequency with which final-/t/ clusters would be tested in this study would be lower than that with which they actually occur in spoken language.

C1	*Total number of tokens (# clusters)*	*C2*	*Total number of tokens (# clusters)*
/b/	7-8 (1)	/t/	27-29 (8-9)
/d/	7 (1)	/n/	10-11
/h/	6-7	/r/	9-10
/s/	6 (5)	/s/	6-7 (2-3)
/ɣ/	5-6 (1-2)	/k/	6 (1)
/k/	5-6 (1-2)	/l/	5-6
/v/	5-6 (1)	/x/	4-5
/p/	5-6 (2)	/m/	3-4
/m/	5	/f/	3-4 (0-1)
/w/	5	/p/	3-4
/t/	5 (1)	/ŋ/	2
/l/	4-5	null	4
/z/	4-5		
/r/	4		
/n/	3-4		
/f/	2		
/j/	1		
null	5		

Table 3.4. Distribution of consonants for the core word list. C1 and C2 refer to word-initial and word-final consonants respectively. The term 'null' means that no consonant was present (i.e., the word began or ended with a vowel).

The second example concerns the phoneme /r/, which has a rhotic pronunciation in word-final position and was observed to be the third most common word-final consonant in the top 1000 words of Dutch. Some instances of word-final /r/ involve inflectional or derivational morphemes. For example, /r/ is the final consonant in the comparative form of most adjectives and it occurs at the end of the adjectival suffix –*baar* (which is approximately equivalent to the English '–able'). However, in contrast to word-final /t/, these morphemes are not extremely common and it is by no means clear that a substantial proportion of the word-final instances of /r/ would have a negligible effect on intelligibility. Furthermore, a Dutch speaker would be able to retrieve a very large number of common, monosyllabic words that end in /r/ reasonably quickly, many of which share English etymology (e.g., /ha:r/ – 'hair' or 'her', /dø:r/ – 'door'). Therefore, the high frequency of

word-final /r/ encountered in the sources was maintained in the present study (see Table 3.4). The frequency of word-initial /r/ was more difficult to choose. Some of the sources showed /r/ to be relatively uncommon in this position (a rating of 1-2), while other sources produced a much higher rating (4 or 5). These differences were partly due to the fact that /r/ rarely occurs at the beginning of function words, causing it to have a low ranking in lists of word-initial phoneme frequency that are based on transcribed speech. However, the discrepancy cannot be solely explained by this factor, as there was also a large difference in frequency between the two lists of common Dutch words.[8] A further possible explanation is variation in the semantic level of the data. Firstly, one of the lowest frequencies of word-initial /r/ was reported in an analysis of child-directed speech (van Severen et al., 2013). Secondly, the word list that produced the highest frequency of /r/ in word-initial position (Memrise, 2014) was examined more closely and was found to contain a reasonably large number of words of a relatively high semantic level (e.g., *reactie* – 'reaction', *rekening* – 'account', *relatie* – 'relationship'). Further research would be required to confirm these proposed explanations. For the present purposes, it was decided that initial-/r/ would be assigned a rating of 4, which in fact corresponded to the frequency calculated using the number of dictionary entries beginning with this phoneme.

As illustrated by the above examples, for most phonemes, there was not perfect agreement between the frequencies measured using different sources. Therefore, in general, it was considered more appropriate to specify a *range* of values for the number of occasions on which the phoneme would be tested (see Table 3.4). The table also specifies how many tokens of each phoneme should be produced as part of a cluster. For example, the voiceless alveolar fricative /s/ rarely appears as a singleton in initial position in Dutch, so the majority of initial-/s/ tokens involved clusters. It was mentioned above that information about the precise phonemic composition of clusters was extracted from the word lists (Dutch 101 and Memrise). The results of this analysis broadly agreed with previous publications (Jongstra, 2003). When constructing the word list, an attempt was made to replicate the frequency distribution of the specific consonant clusters seen in everyday speech. However, due to the multiple constraints at play, this was not always possible. For example, /sl/ was found to be a more common word-initial cluster than /sp/, but the word list included one example of the latter and none of the former. Finally, a small number of words had a CV or a VC syllable shape (denoted by 'null' in Table 3.4). These words were included to allow consonant-addition errors to be detected, following Kent et al. (1989).

[8] As was the case for all of the lists of "common" Dutch words found on the internet, there was almost no information about the sources of the data.

As stated, approximately 15% of the word list was reserved for testing phonemes more often than would be warranted based on their occurrence in the language. Table 3.5 shows the additional phonemes that were tested. The extra tokens of initial /h/ were included to allow Kent et al.'s (1989) 'glottal vs. null' category to be investigated with reasonable reliability. The remaining additional tokens were chosen so that most of the rare phonemes of Dutch would be tested on a reasonable number of occasions (at least three). An exception was the voiceless labiodental fricative, which was only tested twice. This is because it was not possible to come up with three words beginning with /f/ that were sufficiently contrastive. Given that Kent et al.'s (1989) approach focuses on *phonetic-contrast* errors, it was decided that the contrasts that may be tested using /f/ ('fricative place' and 'initial consonant voicing') would be better tested using other phonemes.

Phoneme	*Word position*	*Number of tokens*
/j/	initial	2
/h/	initial	2
/ʃ/	initial	3
/ʏ/	medial	1
/y:/	medial	1-2
/ø:/	medial	2
/ɔu/	medial	1
/œy/	medial	1-2
/ŋ/	final	1

Table 3.5. List of supplementary phonemes.

Having decided on the phonemic distributions of the three segments, the next step was to choose the words themselves. The main criterion was that each word should form a minimal pair with a large number of other words that differ in just one phonetic feature. Table 3.6 shows the Kent et al. (1989) phonetic-contrast categories that were expected to be relevant to speakers of Belgian Dutch.[9] A new category was added (monophthong-diphthong) on the basis that it has been observed in narrow transcription studies (see Chapter 2). The word list is shown in Appendix 3. It was largely designed using the

[9] This table includes the nasal place contrast, despite the fact that this is thought to be difficult to perceive (Narayan, 2008; Black, 1969). Since normative data were being acquired in this project, this would allow examination of the extent to which such confusions arise for perceptual reasons.

categories in Table 3.6, but the minimal-pair possibilities were not confined to these categories, as it was unlikely that they would fully capture the range of articulatory errors to be observed in Dutch speakers. Therefore, the target words were chosen to have as high a neighbourhood density as possible. This approach was particularly important for phonemes that are not part of English phonology. For example, /r/ is produced as a trill in Belgian Dutch, meaning that the phonetic confusions suggested by Kent et al. (1989) for this phoneme (/r/-/l/ and /r/-/w/) may be less likely to occur. Target words that began or ended with /r/ were therefore chosen such that they would allow for a wide variety of reasonably close phonetic confusions, including /l/, /w/, fricatives and plosives. The word list in Appendix 3 also shows the multiple-choice distractors for each token, which were chosen in the light of the orthographic-transcription findings (see Section 3.4.3).

Label	*Phonetic contrast*	*Dutch word-pair example(s)*
1.	Front-back vowels	/beːt/ – /boːt/, /kɛn/ – /kɔn/
2.	High-low vowels	/pɪn/ – /pɛn/, /but/ – /bɔt/, /zin/ – /zɪn/
3.	Long-short vowels	/maːn/ – /mɑn/
4.	Monophthong-diphthong	/bɛt/ – /bɛit/, /bɔut/ – /bɔt/
5.	Voiced-voiceless consonants (syllable-initial)	/poːt/ – /boːt/, /tu/ – /du/, /fɛl/ – /vɛl/
6.	Fricative place of articulation	/veːr/ – /zeːr/, /gɑf/ – /gɑs/
7.	Stop and nasal places of articulation	/ku/ – /tu/, /vɑk/ – /vɑt/, /boːm/ – /boːn/, /mɑt/ – /nɑt/
8.	Stop-fricative	/zɪt/ – /dɪt/, /pɑs/ – /pɑt/
9.	Stop-nasal	/beːn/ – /meːn/
10.	Syllable-initial [h] vs. null	/hɔut/ – /ɔut/
11.	Initial consonant-null	/fits/ – /its/
12.	Final consonant-null	/meːn/ – /meː/
13.	Initial cluster-singleton	/slɑŋ/ – /lɑŋ/
14.	Final cluster-singleton	/rɛist/ – /rɛis/

Table 3.6. Phonetic-contrast categories that are likely to be relevant to Belgian Dutch.

A further criterion for choosing the word list was that a subset of the words (at least 20) should be amenable to unambiguous pictorial representation, to allow future studies to examine the effect of presentation mode (picture naming vs. word reading) on the error rate. The possibility of generating the entire word list to be amenable to pictorial

representation was considered. However, this proved to be too big a constraint. A comparison of the two elicitation modes could have both theoretical and practical value. From a theoretical perspective, one might expect to observe greater accuracy for word reading, due to the presence of orthographic cues. Methodologically, picture naming could perhaps be regarded as a closer approximation to spontaneous speech, in which case it would yield error profiles of greater external validity. Due to the limited availability of listeners, the picture-naming data were not routinely analysed for all speakers in the present study. However, some preliminary findings are reported in Chapter 8.

In the reading task, single words were presented using PowerPoint in the lower-case Calibri font (see Fig. 3.3). Participants were allowed to proceed at their own pace and the author used the shift keys to progress to the next stimulus when the subject appeared to be ready. The word list was separated into four parts; after each set of 30 words, there was a short break during which a scenic picture appeared on the screen, which usually generated a brief conversation. Participants were informed that they were permitted to repeat a stimulus if they were aware that they had made a reading error. The stimuli were presented in the same order for all participants. The possibility was considered of using a counterbalanced design, with two different word orders, to determine whether there was an effect of word order on accuracy (e.g., due to fatigue or accustomisation to the task). However, there were too few participants to justify such a strategy. Therefore, it was decided that the word order, along with other variables such as the task order, would be kept constant in this study, so that between-subject comparisons would have the greatest possible validity. The word order was chosen such that each phoneme appeared with an approximately even distribution throughout the task (e.g., ~25% of words beginning with /d/ occurred within the first set of 30 words, and so on). Thus any temporal effects would be expected to affect all phonemes and phonetic contrasts equally.

The phonemic composition of the subset of words that was also presented pictorially (see Appendix 3) approximately matched that of the core word list. The picture stimuli were tested on neurotypical Belgian speakers, to ensure that they elicited the correct target. As a result of this procedure, a number of items were changed before the picture stimuli were finalised. During the interview, speakers were instructed to name the picture using a monosyllabic word uttered in isolation (i.e., with no definite or indefinite article). They were further informed that if a coloured rectangle (rather than a picture) filled the screen, then they should name the colour. Finally, some of the stimuli required them to name a part of the picture, which was indicated by an arrow (e.g., a person's chin). The practice words included examples of each of these scenarios. During the task itself, if the target word was

not produced, then the author asked the speaker to think of another word to describe the same object. In the majority of cases, repetition of this instruction eventually yielded the desired target. If not, then the author offered semantic cues, followed by phonemic cues, and eventually the target itself (for the speaker to repeat).

Figure 3.3. Example of the visual appearance of the single-word stimuli in (a) picture naming and (b) word reading. The target word was /bo:t/ ('boat').

3.4.3. Multiple-choice distractors

The three multiple-choice distractors chosen for each word are presented in Appendix 3. Note that there are no distractors corresponding to the word /ʃa:l/. This target was omitted from the multiple-choice investigations on the grounds that it did not produce a single error in the free-response mode. In common with Kent et al. (1989), each foil was chosen such that it formed a minimal pair with the target word. However, Kent et al.'s distractors were chosen *theoretically*, based on a list of phonetic contrasts that were thought to be prone to disruption in dysarthria. In the present study, an empirical approach was adopted and the foils were chosen based on the types of phonetic-contrast error observed in the orthographic-transcription study. The following paragraphs describe this methodology.

In the case of consonants, most of the errors observed in the orthographic-transcription study corresponded to one of Kent et al.'s (1989) phonetic-contrast categories. The phoneme /r/ was sometimes transcribed as /l/, but the most common substitution for /r/ was a fricative (and fricatives were also perceived as the rhotic). These substitutions involve a contrast in more than one phonetic feature, meaning that they do not meet the criterion for phonetic-contrast analysis as defined in this thesis. In fact, in the case of confusions between word-initial /r/ and the voiceless velar fricative (which sometimes arose), all three consonant dimensions (voice, place and manner) are affected. Nevertheless, there would have been little point in excluding such substitutions from the

multiple-choice study and replacing them with substitutions that were not observed. Therefore, the foils were chosen so as to allow both /r/ vs. /l/ and /r/ vs. fricative errors.

For vowels, the intention had been to use a similar approach to Kent et al. (1989) and thus to create distractors that differed primarily in just one articulatory dimension, e.g. duration, height, lip-rounding, or backness, with perhaps an additional category to represent monophthong-diphthong confusions. It was presumed that it would be possible to categorise the common vowel substitutions in the orthographic responses according to this schema. However, as reported in Chapter 4, with a few exceptions, the observed vowel confusions did not lend themselves to such simplistic categorisation. Rather, most of the errors involved simultaneous changes in more than one articulatory dimension. For example, a reasonably common error was for the vowel /ɔ/ to be perceived as /ɑ/, which represents a combination of lowering, fronting and unrounding. Large differences in a single articulatory dimension, equivalent to the Kent et al. (1989) front-back contrast category (e.g., English *feed* - *food*), were simply not observed and thus were deemed inappropriate for assessment by a multiple-choice protocol. A consequence of this finding was that, with the exception of the category 'monophthong vs. diphthong', which applies to a wide variety of phoneme pairs, it was not possible to condense the individual vowel confusions observed in the free-response study into a smaller number of categories each based on a single articulatory dimension. Instead, each frequently-observed, *phoneme-specific* vowel confusion (such as /ɔ/ - /ɑ/) was considered as a category in its own right, and the multiple-choice foils were chosen to conform to these categories.

Having decided on the set of consonant and vowel contrast categories to be tested in the multiple-choice study, the following strategies were invoked to choose the distractors for individual words. Firstly, for words that produced a wide range of responses, greater priority was given to errors that were observed in a larger number of speakers. However, this situation was not common. In the majority of cases, it was possible to incorporate the *full range* of errors within the set of three distractors, including errors that occurred in just one or two members of the cohort. There were also target words that only produced one or two error-types in total, meaning that there were "spare" distractors. An example was the word /bɛt/, which was only ever transcribed in two ways – either correctly or as the word /bɪt/. For such targets, the remaining distractors were chosen so as to correspond to a phonetic-contrast error which, while not observed for that particular word, was one of the contrast categories under investigation. While making all of these individual decisions, the *global* constraint that acted on the list of distractors was taken into consideration, namely, the requirement to test all of the contrast-error categories on a "reasonable"

number of occasions. In the case of consonants, the minimum number of occasions on which any given contrast category was tested was 10. However, as explained above, the vowel categories consisted of confusions between specific pairs of phonemes, and it was not possible to test all these confusions on a reasonable number of occasions. Therefore, the reliability with which error rates could be measured for the vowel categories was limited.

3.4.4. Spontaneous speech

The final task in the interview consisted of the production of a spontaneous-speech sample. These stimuli were used to examine the correlation between single-word intelligibility and a quantitative measure of intelligibility derived from spontaneous speech (Lagerberg et al., 2014). The aim was to induce a monologue, rather than to hold a conversation that involved turn-taking. The participants were asked to speak for one or two minutes on a subject of their choice, such as work, hobbies or family. Participants who did not seem to find the task too challenging or tiring were requested to produce two or three monologues on different subject matters. This would allow the within-subject variability in spontaneous-speech intelligibility to be calculated.

3.5. Listening sessions

The data for the listening sessions were prepared using the free, open-source audio editor Audacity®. In the case of word reading and picture naming, each word was saved to a separate sound file. Occasionally, upon listening to a word, it was decided that it had to be discarded due to an unreasonable level of recording noise (e.g., due to clipping) or environmental noise. In one participant (S5), some words were missing due to a recording failure, and in the case of S9, missing words were mainly due to his reading errors.[10]

For spontaneous speech, the starting point of the monologue was the first word uttered by the speaker. The monologues were then divided into utterances that, as far as possible, corresponded to the semantically natural pauses produced by the speaker (so as to avoid interfering with the content). Following Lagerberg et al. (2014), the end-point of the monologue was the end of the utterance that contained the 100th word. Each utterance was

[10] The number of words analysed for each speaker with dysarthria is listed in Chapter 4. There were no missing words for the neurotypical speakers, as they were all interviewed in their homes or places of work where the level of background noise was lower (relative to the hospitals). Furthermore, when background noise was perceived, it was not considered inappropriate to ask these speakers to repeat words, as they were judged to be at low risk of negative effects such as fatigue. This was not always the case for speakers with dysarthria.

saved as a separate sound file. The number of utterances per monologue varied from speaker to speaker, but typically ranged from 12 to 18. Monologues produced by neurotypical speakers were not analysed due to insufficient numbers of listeners.

The listening sessions were conducted using the online survey platform Qualtrics (Qualtrics, Provo, UT). Firstly, two "live" listening sessions were carried out, one at the University of Antwerp and one at the University of Brussels, using postgraduate students studying language sciences. Although the students received their training as a group, they each had their own computer and headphones so that they could work through their assigned Qualtrics session at their own pace. The remaining sessions were performed by listeners who were recruited online using a variety of methods (e.g., personal contacts and site-wide emails distributed among students of the University of Antwerp). These listeners performed the session at home or at their place of work / study and received their training within the Qualtrics platform. The total number of independent listeners was ~90, with some of these individuals performing more than one listening session.[11] In general, the recruitment of listeners and the acquisition of listening data proved to be extremely difficult, for a variety of reasons. Furthermore, many of these difficulties did not come to light until the deadline for acquiring data was drawing to a close.[12] Consequently, as demonstrated in the following paragraphs, there was considerable variability in the listening conditions.

The sessions that were sent to personal contacts were designed to have a duration of ~30-40 minutes. The live listening sessions had a typical duration of 45 minutes. Sessions that were sent to unknown listeners were limited to a duration of approximately 15 minutes. The precise format and composition of the sessions varied according to their duration. However, in general, the 15-minute sessions included only single-word stimuli because of the shorter time required for listeners to complete the training (relative to that required for transcribing a monologue). The longer sessions, in addition to the assessment of single words (either by orthographic transcription or by multiple-choice selection), usually required the transcription of a monologue. Each session typically contained 3-4 parts (called "blocks" in Qualtrics), whereby each block comprised data from a different speaker.

[11] In such cases, it was ensured that the sessions were performed several weeks apart to reduce the possibility of listener familiarity with the speech stimuli.

[12] An example of such a difficulty was the failure of the Qualtrics software in the sense that the programme "hung" and the listener had to abandon the session. This was hypothesised to occur when the listener's network connection was too slow to cope with the loading of the sound files. The problem did not arise when the software was being tested by the author and the author's personal contacts, nor did it arise during the initial listening sessions carried out at the universities.

The set of 3-4 speakers was chosen such that their average intelligibility level was approximately equal to the average intelligibility of the cohort, thereby resulting in equal listener burden in all sessions. The change in duration of the listening sessions (from longer to shorter) was unplanned and arose due to the lack of availability of listeners who could be expected to perform longer sessions. As a result, there was variation across the *speakers* in terms of the way in which their single-word utterances were divided up among different listening sessions; i.e., in the initial, longer sessions, it was possible for the listener to assess the full word list for a given speaker, while in later, shorter sessions, the word list of each speaker had to be divided up among several sessions (and hence among a larger number of listeners). This could have been a confounding factor, as greater exposure to stimuli from a given speaker might have resulted in a learning effect whereby listeners started to recognise the speaker's errors. A given word was never presented on more than two occasions in a listening session, to reduce the risk that listeners would become familiar with the targets. For the same reason, when two blocks of a listening session contained some of the same targets, the block corresponding to the speaker of lower intelligibility was presented first. The order of presentation of the words within a block was randomised. Each session containing a unique set of speaker stimuli (i.e., stimuli that were not presented in any other session) was carried out by between 3 and 5 listeners.

Prior to each listening task, a set of written instructions was provided, along with a set of example stimuli with possible responses. These instructions and training materials are described in further detail in the relevant subsections below. In the live sessions, listeners also benefited from verbal instructions and a live demonstration of how to transcribe spontaneous-speech data. In addition, the author was available to answer questions throughout the live sessions.

Regarding the characteristics of the listeners, the call for participants stipulated that they should be native speakers of Dutch who are familiar with the Antwerp accent and are below the age of 40. The survey included questions to verify these characteristics.[13] In addition, there was a question asking the listener to state whether they had any known hearing difficulties and, if so, to briefly describe them. A small number of listeners above the age of 40 were recruited via personal contacts; these tended to be experts in the field, such as phoneticians and SLTs. As mentioned above, the listeners in the live sessions had some relevant knowledge, although when questioned, they reported that they were not highly

[13] Regarding familiarity with the Antwerp accent, listeners were asked to rate this variable on a five-point scale, where a rating of 1 was described as "I hardly ever hear it" and a rating of 5 corresponded to "It is my own accent or the accent of a close friend or family member".

familiar with dysarthric speech. The online listeners consisted of both experts (SLTs and phoneticians) and individuals with no significant prior knowledge of dysarthric speech.

As can be seen from the above paragraphs, there was considerable variability in all aspects of the listening studies, including the characteristics of the listeners, the format of the sessions and the listening conditions. These factors would have introduced unwanted variability into the data. However, it was possible to reduce the effect of some of these confounding factors using the following strategy. For each speaker, the responses of the individual listeners were inspected to identify any datasets that appeared to be "outlying" with respect to either the *number* or the *types* of perceived errors. In such cases, the dataset for the listener in question was discarded. In all but one of these cases,[14] the outlying dataset showed a substantially *higher* error rate than the remaining datasets. Interestingly, there did not always appear to be an obvious reason for the inferior data of the individual in question (e.g., higher age or lower rating of familiarity with the Antwerp accent). Furthermore, some of the rejected datasets originated from the live listening sessions, implying that outlying results were not always due to inferior equipment or unfavourable listening conditions (factors that might play a role in sessions carried out in people's homes). It is possible that these individuals had an undiagnosed hearing problem. McHenry (2011) reported similar outliers in her study. For example, for a speaker with a mild flaccid dysarthria who achieved a mean intelligibility score of 89% in Kent et al.'s (1989) multiple-choice test (an average of the scores across 67 listeners), the lowest score for any given listener was 59%. In common with the present study, information on hearing ability was self-reported.

The remaining subsections (below) describe the written instructions that were provided to listeners within the Qualtrics platform. For most tasks, preliminary training exercises were also offered.

3.5.1. The assessment of single words

In this task, listeners were instructed to listen to the word on a maximum of two occasions (following Whitehill & Ciocca, 2000b) and then to provide their response – either orthographic transcription or selection of a multiple-choice option. The possibility was considered of configuring the software such that it only played the word twice, after which the sound file would no longer be accessible. This would prevent listeners from exceeding the maximum number of allowed listening occasions. However, such an approach would

[14] The dataset that was rejected due to an unusually *low* error rate belonged to an SLT with considerable experience in listening to dysarthric speech.

run the risk that listeners do not hear the stimulus at all, or that they only hear it once – for example, if an unexpected source of background noise were to arise during one or both of the listening occasions.

In the orthographic-transcription task, listeners typed their response in an open field. The listeners were informed that all words constituted real words of Dutch containing one syllable. Furthermore, they were told that the word could be common or uncommon, could involve any part of speech (e.g., noun, verb), and could be a proper noun (e.g., a name or place). Written examples of words meeting these various criteria (not drawn from the word list) were provided. It was emphasised that if the word was not intelligible, they should type the real word of Dutch with the closest similarity to what they perceived. Thus, the field should not be left empty and they should refrain from typing multisyllabic words or 'words' with no meaning. Having given these instructions, the programme then provided some example stimuli with plausible responses. It was explained to the listener that there is no right answer and that they should not be concerned if their perception differs from the suggested response. The example stimuli consisted of tokens of both low and high intelligibility. Cases of epenthesis (e.g., schwa insertion in a consonant cluster) were included. In such productions, the word often sounded as though it contained two syllables. Thus the purpose of these examples was to emphasise to the listener that the transcribed word should always be monosyllabic. Despite all these precautions, there were cases of missing and invalid responses (e.g., pseudo-words, monosyllabic words), in which case the response was discarded. These stimuli were offered again in a different listening session to ensure that all words uttered by the speaker were assessed by the same number of listeners. Missing or invalid responses usually arose for words that were highly unintelligible and hence difficult to perceive as any real word of Dutch. This is undoubtedly a disadvantage of orthographic-transcription protocols and may explain why clinical dysarthria assessments usually recommend a forced-choice protocol for severe speakers.

Due to the straightforward nature of the multiple-choice task, the instructions for this part of the session were very brief. The listeners were asked to ensure that they studied all four options before making their choice. If the word seemed to be completely unintelligible, then they were instructed to choose an option at random rather than leaving the item unanswered. No example stimuli were provided for the multiple-choice task. This is because all sessions that included this task had already required the listener to assess single words by means of orthographic transcription. Thus the listeners were already familiar with the nature of the stimuli.

3.5.2. The assessment of spontaneous speech

Spontaneous-speech intelligibility was assessed using the method proposed by Lagerberg et al. (2014). According to this approach, listeners use orthography to record every word that they can understand, and in the remaining portions of speech (which are unintelligible), they count and record the number of syllables. The intelligibility metric is then calculated as the number of syllables in the transcribed words as a percentage of the total number of syllables perceived in the monologue. Lagerberg et al. (2014) provided a detailed description of how to train listeners to carry out this task, including a practice session in which specific feedback was given to listeners regarding their transcriptions. Unfortunately, it was not possible to replicate this training method in the present study. Firstly, approximately half of the monologue transcriptions were carried out online. Secondly, even in the case of live sessions, there was insufficient time to administer a prolonged period of training. For the online sessions, instructions and training exercises were provided within the Qualtrics programme, as described in the following paragraphs. Listeners in the live sessions also benefited from a live demonstration of the transcription technique and the opportunity to ask questions at any time. Since the live listening sessions were carried out first, it was possible to refine the online instructions in the light of any questions that had arisen. Furthermore, the online protocol was tested and further refined through preliminary listening sessions carried out by personal contacts of the author.

Listeners were informed that they were allowed play each utterance on a maximum of three occasions. As with single words, their adherence to this instruction was left as a matter of trust. The number of listening occasions was one greater than that allowed by Lagerberg et al. (2014), as pilot listening tests showed that, for speakers of low intelligibility, it was not possible to meet the transcription aims based on just two listening occasions. Unlike single-word transcription, where the sole objective was to match the perceived signal to a known word of Dutch, the transcription of spontaneous speech included two elements: word identification and syllable counting. The burden of achieving these tasks after just two listening occasions was found to be too great, at least in some of the monologues. The fact that Lagerberg et al. (2014) did not report the same finding suggests that the utterances in their study (produced by children with speech delay) were less challenging to the listener than the dysarthric utterances of the present study. Further evidence in support of this statement is that Lagerberg et al. (2014) measured utterance length in terms of the number of *words*, whereas in this study, the metric used was the number of *syllables*. This is because it was not always possible to identify word boundaries.

The transcription instructions were presented at the beginning of the task. An abridged version of these instructions also appeared at the bottom of the screen throughout the task, so that listeners did not have to rely on their memory. The full set of instructions, translated into English, was as follows:

"You may listen to each utterance three times. Type all the words that you can understand. If you can determine a word from the context, then that word may be typed, even if the word itself is not intelligible. If, however, you cannot identify a word with a reasonable degree of certainty, then **do not guess**. Instead of typing the word, denote every syllable of the word with the symbol '0'.

You should type all of the following:

- words such as "bang" or "splat"
- whole words that the speaker repeats
- proper nouns or foreign words (if you can identify them with certainty; otherwise, note every syllable with '0')
- dialect (please replace with standard Dutch words that carry the same meaning)
- numbers (please write them in full, e.g., "forty six" and not "46")

You do <u>not</u> need to type:

- words without any meaning (e.g., "um", "er")
- parts of words

To assist you in understanding the monologue, the topic of conversation is provided."

The above instructions largely replicated those that were developed by Lagerberg et al. (2014). An exception was the provision of the topic of conversation, which was a new element introduced in the present study. Typical monologue titles included "My first job", "My hobby" and "Holidays". The rationale for providing this information was that in a real-world communicative situation, where a speaker is relating a personal narrative, this does not take place in a vacuum. Rather, the speaker is normally prompted by a question from their conversation partner or is contributing to the existing topic of conversation among a group of people. Further discussion of the advantages and disadvantages of providing the listener with the context is provided in Chapter 8.

Having been provided with the above instructions, the listener was presented with two demonstrations of the technique. The first demonstration involved 5 utterances produced by a speaker of reasonably high intelligibility. The second demonstration involved 6 utterances produced by a different speaker of lower intelligibility. In both cases, the topic of conversation was provided. Beneath each utterance, there was a possible transcription for the listener to study. An example of such a transcription was as follows:

en ik ga veel met de buurvrouw 000 (English: “and I go a lot with the neighbour 000”).
As was also the case for single words, it was emphasised that there is no right answer and that the listener should not be concerned if their own transcription would have differed from that suggested. Following the demonstrations, the listener was given the opportunity to carry out their own practice transcription of a third monologue. The monologue consisted of three utterances and had a moderate-high level of intelligibility. No feedback was given on this task (e.g., the author’s own “suggested” transcriptions), as the listener might have viewed these transcriptions as the right answer and, if there were significant discrepancies, might have doubted their ability to carry out the task.

As mentioned, the instructions and training procedures described above were developed and refined using pilot tests carried out online by close personal contacts of the author. The later iterations of these tests showed considerable promise for acquiring monologue transcriptions and intelligibility measures of a reasonably high level of inter-listener agreement. However, some loss of quality was expected in the final study, as acquaintances and strangers cannot be expected to have the same level of motivation as close personal contacts. Therefore, they may not be willing or able to devote sufficient time to reading the instructions or following the training exercises. As described in Chapter 7, it was indeed discovered that the final data were subject to two major deficiencies: (a) low levels of inter-listener agreement regarding the number of syllables in the unintelligible portions, and (b) occasional guesswork in the portions that were considered intelligible (i.e., transcriptions that were clearly based on a low level of certainty). Fortunately, as described in Chapter 7, it was possible to develop techniques to correct for both of these deficiencies, resulting in intelligibility measures that were deemed to be of high accuracy.

4. Study 1: Orthographic transcription of single words in speakers with dysarthria

4.1. Aims

The main aim of this study was to determine whether it is justified to confine the errors observed in Belgian Dutch speakers with dysarthria to a reasonable number of phonetic-contrast categories. As explained in Chapter 2, this question was addressed using an exploratory approach because it is multi-faceted and difficult to subject to quantitative testing given (a) the current state of knowledge on the matter and (b) the limited resources of the project (in particular, the low sample size and the use of a novel, untested word list). The second aim of Study 1 was to obtain preliminary information regarding the phonemic and phonetic-contrast errors of Belgian Dutch speakers with dysarthria. Due to the lack of published data on perceptual, articulatory errors in this population, the second question was likewise exploratory in nature, with a particular interest being the types of substitution observed for phonemes that are not part of English phonology (e.g., velar fricatives). The third aim, also exploratory, was to contribute to the evidence base for the inter-rater reliability of the identification of phonetic-contrast errors via orthographic transcription.

4.2. Data analysis methods

The first analysis conducted in this study involved calculating metrics of overall speaker *intelligibility* (i.e., accuracy). This was implemented both at the whole-word level and for each individual word segment (i.e., C1, V and C2). For each speaker, each of these four accuracy metrics (word, C1, V and C2) was obtained by considering all observations, from all listeners, as one dataset. The metric was then calculated as the number of words (or segments) transcribed correctly as a percentage of the total number of words (or segments) uttered by the speaker. In the case of segmental accuracies, segments that were present in the target word, but perceived as omissions, were included in the denominator. The reason for calculating summary measures in this way (i.e., by pooling the data over all listeners) was that the number of *independent* listeners who judged each speaker varied, depending on how the speaker's single-word utterances were divided up among different listening sessions (see Chapter 3). This meant that it was not possible to determine a mean accuracy metric across listeners, nor a measure of variance (e.g., standard deviation), that would have a consistent meaning in all speakers. As for the number of listeners assigned to each word, this varied from a minimum of three to a maximum of five. Initially, three listeners judged each word uttered by each speaker. After analysis of these data, the

speakers who seemed to yield the greatest levels of inter-listener variability (in terms of word accuracy) were assigned additional listeners, up to a maximum of five. This strategy maximised the utility of the available listeners. A repeated-measures ANOVA was used to determine whether the mean accuracies of the three word segments (C1, V and C2) were significantly different. This was followed by Bonferroni-corrected pairwise comparisons.

The second analysis related to the accuracies with which specific *phonemes* were perceived. This analysis was carried out just for the consonant phonemes, as vowel articulation is less variable in terms of the range of articulatory gestures. Therefore, in comparison with consonants, where each phoneme differs in terms of its combination of the features voice, place and manner, in the case of vowels, the relative accuracies of specific phonemes are of lesser interest.[1] The outcome measure used to reflect the vulnerability of consonant phonemes was the error rate, which was the number of incorrect transcriptions relative to the number of occasions on which the phoneme was uttered by the speaker. The error rates were visually inspected to assess the relative vulnerabilities of all the consonant phonemes of Dutch at each word position (C1 and C2). Linear mixed models were used to determine, for each word position, whether there was an effect of (a) manner of articulation and (b) place of articulation on the error rate. It was not possible to examine the interaction between place and manner due to rank deficiency of the data.

The third analysis was the classification and quantification of errors in terms of *phonetic-contrast* confusions. To help the reader understand the significance of this analysis, it is worth briefly repeating some of the discussion in Chapter 2 regarding outcome measures. In a free-response mode, the researcher has no control over the words that may be substituted for a given target. Therefore, the perceived contrast errors are influenced by functional load. For example, if it were to be determined that 'stop place' errors arise more frequently than 'stop vs. nasal' errors, then this could, at least in part, be due to the existence of a greater number of minimal pairs that are based on stop place of articulation than on the stop-nasal contrast. Therefore an error that arises frequently may not be strongly indicative of speech production (or perception) difficulties. This has important implications for comparing contrast errors among studies. A meaningful comparison with the present study is only possible when the other study also uses an open-response mode (so that neither study controls for functional load). Further, since the functional loads of different phonetic contrasts are language-dependent, the comparison would need to be made within the same language or with a language that has a very similar phonetic-contrast

[1] There may be interesting differences in the error rates for different *categories* of vowel, e.g., high vs. low vowels, but this would come to light in the analysis of phonetic-contrast errors.

distribution. To the best of the author's knowledge, there are no studies that meet these criteria, meaning that the present findings cannot be directly compared with previous data on phonetic-contrast error profiles in dysarthric speakers. Accordingly, a thorough discussion of phonetic-contrast errors from the perspective of impaired *speech production* is deferred to later chapters, in the light of the information obtained from the normal-control and multiple-choice studies about perceptual distinctiveness and functional load. Thus, the analysis of the phonetic-contrast data in the present chapter is relatively limited. Firstly, it involves computing the relative frequency of each contrast error, which is a measure of how often the error arises as a proportion of the total number of errors yielded by the speaker at the same word position (C1, V or C2). The second analysis procedure involves defining a meaningful set of phonetic-contrast categories to be used in Dutch dysarthria assessments. The methods used to achieve this are intricate and are best described in the context of the data. Thirdly, the proportion of phonemic-substitution errors that can be captured by these phonetic-contrast categories is reported. In defining these categories, the Kent et al. (1989) nomenclature is used whereby each category is named after the phonemic substitution *reported by the listener*. As explained in Chapter 2 (Section 2.1.5), this approach has the advantage of describing the relationship between the target and the transcribed phoneme (or phoneme group) in an unambiguous way.

The fourth analysis sheds light on the inter-rater reliability of the identification of phonetic-contrast errors via orthographic transcription. The level of inter-observer agreement should ideally be calculated for the *outcome measure* – in the present case, the profile of phonetic-contrast errors. However, this was not possible using the current data, as most listeners did not transcribe the full set of words for a given speaker. Therefore, a method was devised that involved calculating, for each erroneous word, the ratio of the number of non-unique phonetic-contrast errors (i.e., errors that were transcribed by at least two listeners) to the total number of errors perceived for that word across all listeners. This metric was then averaged across all erroneous words to produce a measure of "consistency" for the speaker. Previous studies have shown that the higher the intelligibility of a speaker, the higher the level of inter-rater agreement on the nature of the articulatory errors (see Section 2.1.5). To investigate whether there is any such relationship for the present study, Pearson's *r* was used to calculate the correlation between the consistency measure and word accuracy (having first established that both variables pass the Shapiro-Wilk test for normality). The consistency measure in the present study is conceptually different from inter-rater reliability metrics in previous studies; specifically, it is only calculated for erroneous words and does not include agreement on correct items. Therefore, there was no expectation that a positive correlation would be observed.

Finally, it was mentioned in Section 4.1 that a potential limitation of the present study is that the word list had not previously been tested in dysarthric speakers. A rigorous validation process would have been beyond the scope of the thesis and would have required a much larger population sample. Nevertheless, it was considered worthwhile investigating one of the key parameters that might act as a confounding factor: lexical frequency. The frequency of each target word was rated on a 5-point scale based on an online word-frequency database for spoken Dutch (Corpus Gesproken Nederlands). Since word frequency was an ordinal variable, the correlation between frequency and accuracy (i.e., the percentage of correct transcriptions for each target word) was computed using Spearman's rank correlation coefficient.

4.3. Results

4.3.1. Word accuracy and segmental accuracies

Table 4.1 presents word and segmental accuracies for all speakers. Words that lacked an initial consonant (VC words) were excluded from the C1 accuracy calculation, despite the fact that an error at word-initial position was possible ('initial consonant addition'). It was reasoned that such an error did not constitute a consonant error *per se*, as there was no consonant to aim for. Similarly, words with no final consonant were excluded from the calculation of C2 accuracy. The data are also presented graphically in Fig. 4.1. In addition to showing accuracy as a function of word position, the graph plots the *average* phoneme accuracy for each speaker (i.e., calculated across all word positions). Speakers are plotted in order of increasing average phoneme accuracy.

Figure 4.1 reveals that for the six mildest speakers (right-hand side of the graph), the C2 position yielded the highest accuracy, while C1 and V (which showed similar accuracies) were more prone to error. Some of the more severe speakers showed different patterns. For example, Speaker 5 yielded the lowest accuracy at C2 position. He was also the only speaker for whom vowels were the least affected (alongside word-initial consonants; his C1 and V accuracies were almost identical). Therefore, while there is a discernible pattern in the relative vulnerability of the three phoneme positions across the whole cohort, it seems that certain individuals (particularly the more severe speakers) depart from this trend. Averaging across the cohort (see Table 4.1), the highest accuracy was for word-final consonants, while the lowest accuracy occurred in vowels (C2 > C1 > V). A repeated-measures ANOVA showed that the mean accuracies of the segments were significantly different: $F(2,18) = 6.42$, $p = 0.008$. However, post-hoc pairwise comparison tests using the Bonferroni correction revealed that the only significant difference was between V and C2:

mean difference 6.36%, p = 0.02 (two-tailed), 95% confidence intervals for the difference [1.02%, 11.70%]. Further research is needed to differentiate between C1 and C2. However, if it could be shown that the greatest accuracy occurs at C2 position, then this would be unsurprising. In Dutch, the number of possible phonemes is lowest at C2 position, largely due to the devoicing of word-final consonants. Thus, the potential for minimal pairs involving C2 is lower than for the other two word segments.

ID (M/F)	*Diagnosis*	*# words*	*Word accuracy (%)*	*C1 accuracy (%)*	*V accuracy (%)*	*C2 accuracy (%)*
1 (F)	ALS	101	78.0	89.7	**86.9**	95.1
2 (F)	CVA (suspected to be in brainstem)	107	77.3	90.9	**87.2**	94.7
3 (M)	Medulloblastoma / surgical damage (left cerebellum)	107	73.2	86.6	**84.7**	94.4
4 (F)	Surgical damage (fourth ventricle, posterior fossa)	115	49.1	78.8	**71.6**	74.7
5 (M)	Progressive cerebellar atrophy	88	70.6	90.4	89.6	**83.9**
6 (M)	CVA (right cerebellum)	117	83.9	**91.7**	93.7	98.2
7 (F)	CVA (pons / left cerebral peduncle)	115	88.4	96.4	**92.4**	98.4
8 (M)	Cortical watershed CVA (PCA / MCA)	117	68.8	85.5	**80.2**	95.4
9 (M)	CVA (left cerebellum)	95	62.0	**75.1**	81. 9	94.0
10 (M)	Suspected ALS	117	90.7	96.7	**95.2**	98.2
Mean accuracy values (± 1 SD)			74.2 ± 12.6	88.2 ± 7.0	86.3 ± 7.1	92.7 ± 7.6

Table 4.1. Word and segmental accuracies for all dysarthric speakers. For the segmental accuracies, the value in bold shows which segment (C1, V or C2) was most prone to error. Column 3 shows how many words were analysed for each speaker (where the full word list consisted of 117 words).

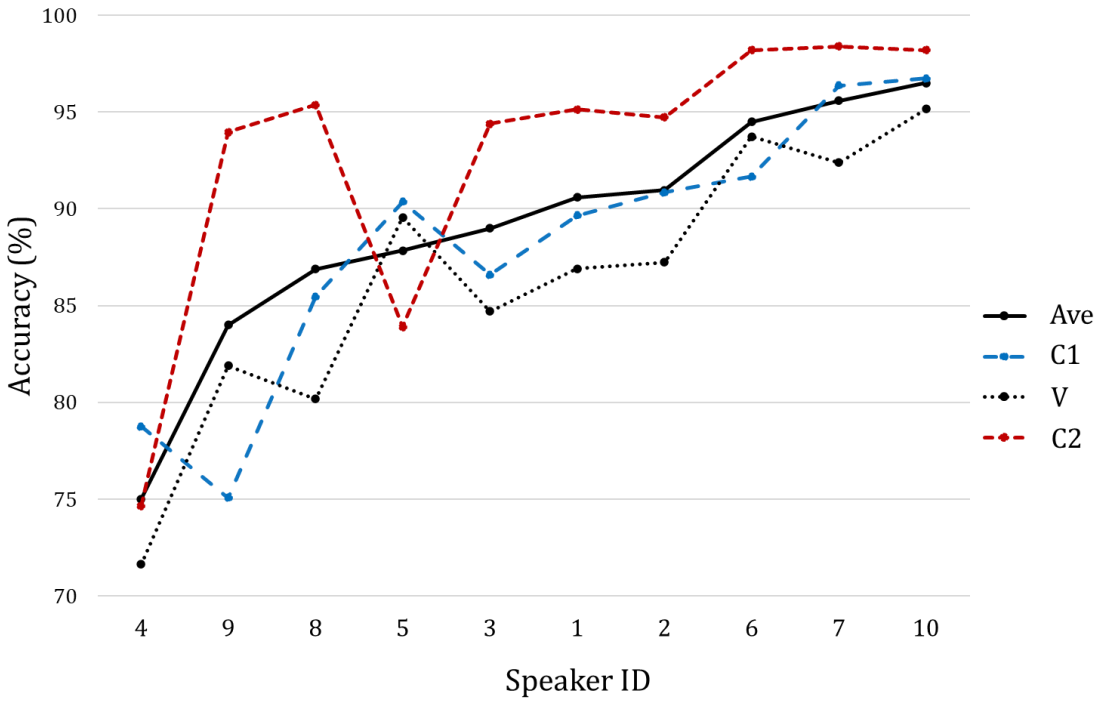


Figure 4.1. Average phoneme accuracy and individual phoneme accuracies for each speaker.

4.3.2. Consonant accuracies

Table 4.2 summarises the mean error rates (averaged across all speakers) for all C1 and C2 consonants. It can be observed that for some consonants, the error rate is relatively independent of position. For example, /k/ and /p/ yield very few errors whether they occur initially or finally. Other consonants show an interaction with position; for example, /f/ is the most error-prone consonant in word-initial position, but it is reasonably accurate in word-final position. It was initially expected that phonemic error rates might be strongly determined by linguistic factors. Specifically, words that contain the more common phonemes of a language might be expected to have a larger neighbourhood density than words containing rarer phonemes, which would then result in higher error rates. This would imply that phoneme error rates are not indicative of articulatory difficulty. Fortunately, there was evidence to suggest that the effect of functional load was negligible. For example, it was frequently observed that when a given phoneme was challenging for the speaker, but did not give rise to a minimally contrastive word based on the types of substitution typically observed for that speaker, the word was perceived as a more distant substitution so that the error would still be accommodated. An example was the word /va:x/ transcribed as /pa:r/ in a speaker who made frequent /v/-/p/ substitutions, because there is no Dutch word [pa:x].[2] Further evidence can be seen from the fact that the

[2] Of course, it is possible that the second confusion, where /r/ was transcribed as [x], was actually perceived by the listener as a clear substitution. However, from examining several of these sorts of instances, it was found that the second error was often not characteristic of the speaker.

highest proportional error for the set of phonemes occurring at C1 position was observed for /f/, despite the facts that (a) this is one of the least common word-initial phonemes in the Dutch language and (b) Dutch words beginning with /f/ have relatively few minimal pairs. Therefore, the assumption that rarer phonemes would produce fewer errors was not borne out, and the error rates reported for individual phonemes in Table 4.2 are likely to reflect the phoneme's articulatory difficulty and/or perceptual distinctiveness.[3]

Word-initial consonant	*Error rate (%)*	*Word-final consonant*	*Error rate (%)*
/f/	25.6	/ŋ/	38.7
/ɣ/	23.2	/m/	20.8
/v/	20.5	/n/	10.3
clusters	15.8	/r/	8.1
/m/	15.1	/f/	8.0
/h/	12.5	clusters	7.4
/l/	11.8	/l/	6.7
/d/	11.5	/s/	4.9
/b/	11.4	/t/	4.6
/t/, /j/	11.3	/x/, /k/	2.8
/r/	9.8	/p/	0.5
/s/†	9.6		
/ʃ/	9.1		
/w/	6.1		
/p/	5.7		
/n/	4.6		
/z/	2.1		
/k/	1.7		

† There was only one word beginning with singleton /s/ in the word list, as this phoneme rarely appears as a singleton in C1 position in Dutch. Therefore the result obtained for this phoneme cannot be considered reliable (although the fact that /z/ was also produced with high accuracy lends support to the finding, as does the relatively high accuracy of /s/ in word-final position).

Table 4.2. Mean consonant error rates across the speakers with dysarthria, separated according to word position (initial or final).

[3] There could be other factors, including a bias in the opposite direction to that hypothesised above. In other words, it could be the case that less common phonemes are *more* likely to be heard incorrectly because the listener has a lower prior expectation of hearing them (Green & Swets, 1966). However, the error rates for /v/ and /f/ were similar (where /v/ is a reasonably common C1 phoneme), suggesting that phoneme frequency was probably not a strong determinant of error rate.

The data in Table 4.2 were also displayed graphically (see Fig. 4.2) in a way that facilitates interpretation according to articulatory theory. Thus, phonemes with an identical manner (left-hand graphs) or place (right-hand graphs) were coded with the same colour. It was assumed that in the Antwerp accent, the speaker attempts to realise /r/ as an alveolar trill (Verhoeven, 2005). From Fig. 4.2a, it can be seen that there is no consistent dependence of C1 error rate on manner, although the three highest error rates occur for fricatives. A linear mixed effects (LME) model with 'speaker' as the random effect and 'manner' as the fixed effect, followed by an *F*-test of the significance of the fixed-effect term, was calculated using the MATLAB function *fitlme*, with the option of the restricted maximum likelihood fit method. Hypothesis testing was implemented using the MATLAB function *anova* with the Satterthwaite correction applied to the denominator degrees of freedom. These procedures confirmed that manner is not significant. When the C1 error rates are categorised according to place of articulation (Fig. 4.2b), the labiodental place appears to be the most vulnerable. However, the overall effect of place is non-significant according to the aforementioned statistical procedure (LME modelling followed by hypothesis testing).[4] The effects of place and manner on error rates for C1 phonemes may also be examined in Fig. 4.3, which plots summary measures produced from Figs. 4.2a and 4.2b.

For word-final consonants, there is a clear effect of manner (Fig. 4.2c), with nasals being the most vulnerable consonant type and plosives showing almost no errors. The LME model followed by hypothesis testing showed that the effect of manner was significant ($F = 8.73$, $p < 0.001$). Post-hoc pairwise comparison tests were conducted using the MATLAB function *coefTest*, with a Satterthwaite approximation. This prodcedure demonstrated that nasals were significantly different from each of the other manners of articulation ($p < 0.01$),[5] while the remaining pairwise comparisons were all non-significant. Figure 4.2d shows that there is no clear evidence of a place effect for word-final consonants, and this was confirmed statistically.

[4] The effect of the level 'labiodental' (where the reference manner level was 'alveolar') was of borderline significance ($p = 0.06$ using the Satterthwaite approximation).

[5] Since there are 5 manner levels at C2 position, there were 10 pairwise comparisons. Therefore, these *p*-values could be Bonferroni-adjusted by multiplying by 10, in which case they would still be significant at less than the 10% level.

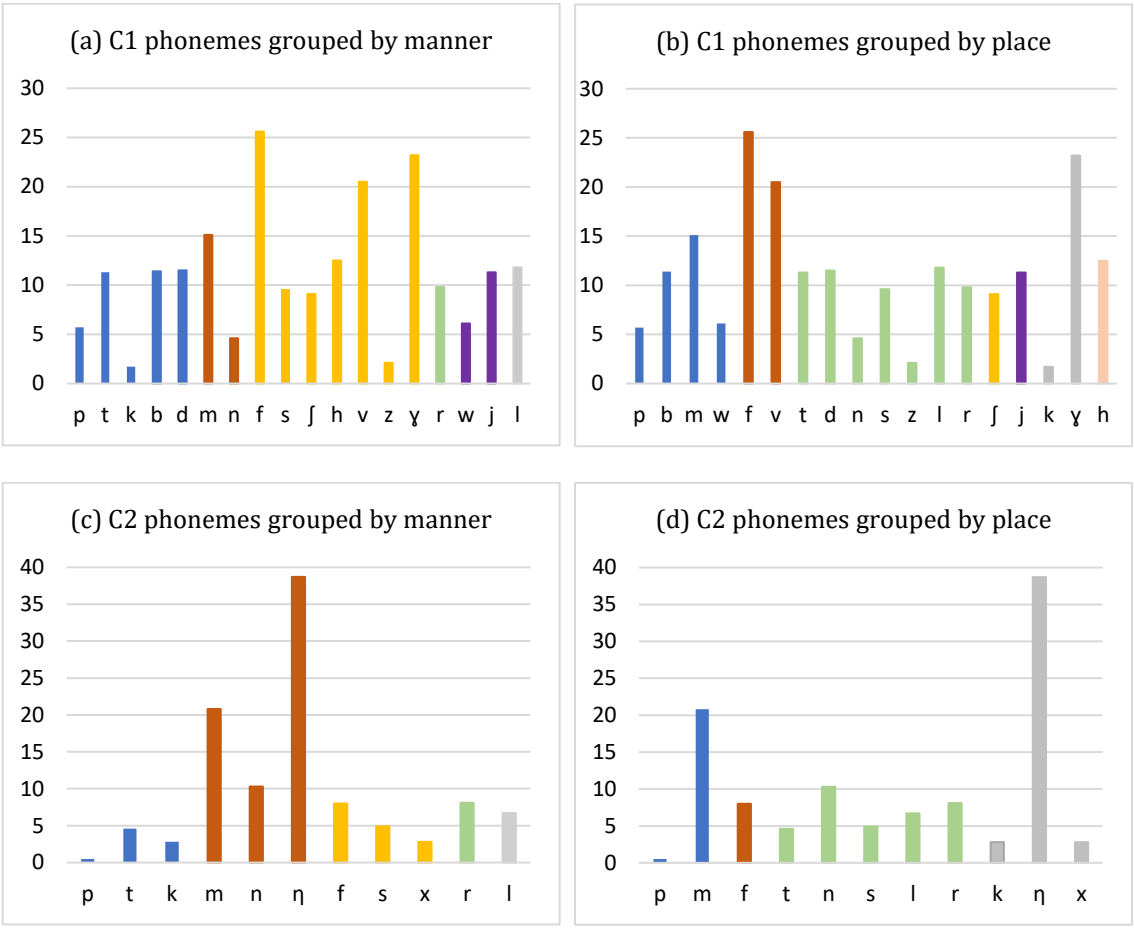


Figure 4.2. Mean percentage error rates across all speakers for C1 (top) and C2 (bottom) phonemes, colour-coded and grouped according to manner (left) and place (right) of articulation.

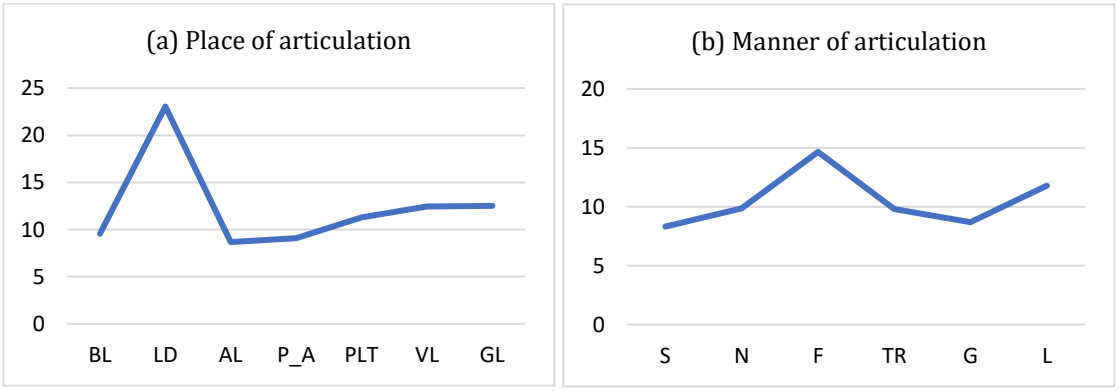


Figure 4.3. Mean C1 percentage error rates as a function of (a) place: BL = bilabial (/p, b, m, w/), LD = labiodental (/f, v/), AL = alveolar (/t, d, n, s, z, l, r/), P_A = post-alveolar (/ʃ/), PLT = palatal (/j/), VL = velar (/k, ɣ/), and GL = glottal (/h/), and (b) manner: S = stop (/p, b, t, d, k/), N = nasal (/m, n/), F = fricative (/f, v, s, z, ʃ, ɣ, h/), TR = trill (/r/), G = glide (/w, j/), L = liquid (/l/).

Figure 4.4 compares consonant error distributions for four speakers. Only the C1 data are shown due to the limited range of consonants at word-final position. It can be seen that although some phonemes appear to be a persistent problem (e.g., /f/), in general, the error

profiles are highly individualised and no obvious pattern emerges. For example, it might have been expected that milder speakers would only produce errors on consonants that, in general, are problematic for the whole cohort, and that the inventory of vulnerable phonemes would grow as the data from more severe speakers are added. However, Fig. 4.4 shows that Speaker 6 (the most intelligible of the four speakers) mainly exhibited errors on alveolar plosives, despite the fact that this was a relatively stable consonant group for many other participants (and especially for other speakers with mild dysarthria). Thus it seems that Speaker 6 had specific difficulties with the alveloar plosives. Inspection of his data revealed that the errors perceived for /d/ were mainly devoicing errors, while all the errors transcribed for /t/ were place errors. Similarly, Speaker 2's dysarthria seems to be characterised by errors on velar consonants, while for other speakers (even Speaker 4, who is the least intelligible), this consonant class is relatively robust.

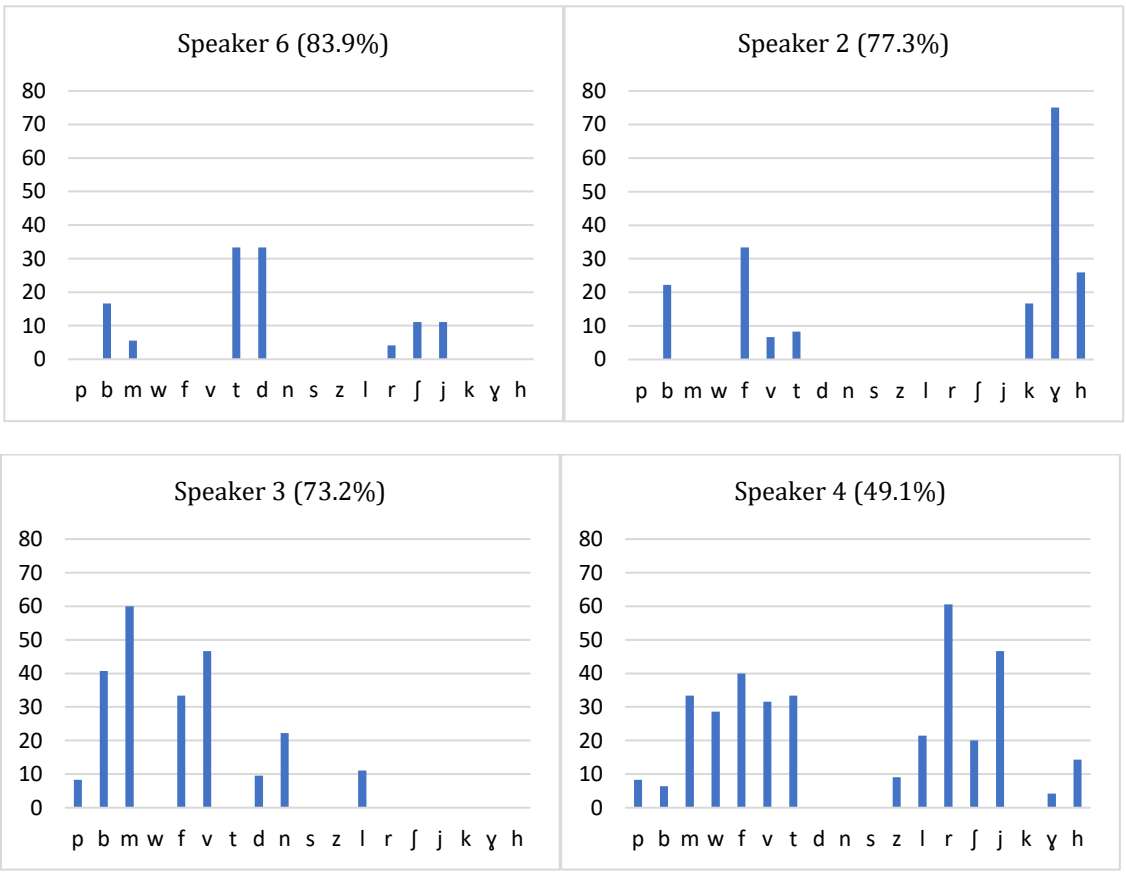


Figure 4.4. C1 percentage error rates for four speakers of different intelligibility levels (shown in brackets as word accuracy). The consonants are organised in terms of place of articulation.

4.3.3. Consonant contrast errors

In this section, the consonant substitutions are analysed with respect to Kent et al.'s (1989) proposed framework. Before presenting the results of this analysis, the issue of inter-rater reliability needs to be mentioned. As explained in Section 4.2, it was not possible to

calculate the level of inter-observer agreement for the *outcome measure* – the profile of phonetic-contrast errors. Therefore, a method was devised that involved calculating, for each erroneous word, the ratio of the number of phonetic-contrast errors that were transcribed by at least two listeners to the total number of errors perceived for that word. A solid understanding of this metric can only be achieved in the light of the findings of the current and following subsections (on phonetic-contrast confusions). Therefore, the inter-rater reliability data are described in detail thereafter (in Section 4.3.5). However, it is worth briefly summarising the main findings at this juncture, so that the reader has some appreciation of the reliability of the data that are about to be presented. The percentage of phonetic-contrast errors (e.g., initial-stop devoicing) that were perceived by at least two listeners ranged from 40.5% to 69.8% across the cohort, with a mean (± 1 SD) of 61.0% ± 9.5%. This means that, on average, almost 40% of the contrast errors transcribed for *any given word* were unique. However, as mentioned above, the inter-rater agreement at the level of individual words does not provide direct information on the outcome measure of interest, which is the total number of errors observed for each contrast category. In other words, two listeners could perceive a similar number of errors for a given category, but distributed differently among the relevant targets. Nevertheless, owing to the relatively low level of inter-rater agreement for each target, the phonetic-contrast data in this thesis are interpreted with caution, and conclusions are only drawn when, in the author's estimation, they are highly unlikely to be peculiar to the current set of listeners. In particular, the findings for *individual* speakers are not discussed in any detail unless based on a large number of perceived errors. Finally, it is worth pointing out that the consistency metric derived in the present study is conceptually different from conventional inter-rater reliability metrics in the sense that it is only calculated for erroneous words and does not include agreement on correct items. Therefore, the reader is cautioned against comparing the average consistency level (61.0%) with measures of inter-rater agreement that are typically quoted in other transcription studies.

The first question that needs to be addressed with regard to consonant contrasts is whether the majority of errors perceived in the cohort can be described using a reasonable number of phonetic-contrast categories. When an error could not be coded, this was either because it represented a contrast category that fell outside the predetermined set of categories or because it spanned more than one category simultaneously (e.g., /f/ → /k/, which involves backing and fricative stopping). As might be expected, the proportion of consonant errors that could not be coded increased with speaker severity. However, even for the three most

severe speakers, the proportion was ≲ 22%.[6] Therefore, for the present population, the concept of describing consonant errors in terms of a reasonable number of phonetic-contrast categories is broadly[7] applicable. Despite the low incidence of "multiple" phonetic-contrast errors (e.g., /f/ → /k), it was decided that rather than leave them out of the analysis, they would be coded by marking an error for each of the relevant categories. In some cases, there was an element of choice as to how those categories could be defined. For example, if the word-initial phoneme /r/ were perceived as /vl/, then this could be coded either as an /r/-fricative substitution plus an 'initial singleton vs. cluster' error or an /r/-/l/ substitution plus an 'initial singleton vs. cluster' error. In such situations, the potential coding categories were compared with the predominant error types perceived in that speaker. The error was then coded such that it best matched the speaker's existing error profile. For target words that contained consonant clusters, errors were coded separately for each phoneme. Thus /br/ → /pl/ was coded as a devoiced stop plus an /r/→ /l/ substitution. Note that substitutions of this kind were not considered to be "multiple errors" because each of the two substitutions could be coded by one error category. Therefore such instances did not contribute to the aforementioned result of 22%.

Table 4.3 shows the most common C1 contrast errors. As explained in Chapter 2, *absolute* error rates (i.e., the ratio of observed errors to potential errors) cannot be calculated for phonetic contrasts, because the denominator is unknown. The proportional errors shown in Table 4.3, therefore, have a different meaning. They represent the number of errors within a particular contrast category (e.g., 'stop place') relative to the *total number of C1 errors* yielded by the speaker. This normalisation process was found to be necessary in order to avoid the situation where error values averaged across the cohort were dominated by the results of one or two speakers. Furthermore, it is a useful metric when comparing individual speakers, as it provides information about the relative importance of a given contrast error compared to all other contrast errors for that speaker.

[6] The reason for citing a threshold at this stage, rather than an exact proportion, is that the latter would require the definitive set of contrast categories to have been chosen. This is an intricate process and is more appropriately addressed in the Discussion section (see Section 4.4.1). An example of one of the decisions that needed to be made was whether, for some of the contrast categories, it would be justified to allow voicing errors alongside the confusion of interest (i.e., a place or a manner contrast), as was the case in previous studies (Kent et al., 1989; Haley et al., 2000). In Section 4.4.1, the exact proportion of errors that were amenable to categorisation is cited.

[7] As stated in the previous footnote, in common with prior studies, not all of the categories met the strict definition of a contrast in a single phonetic feature. This was particularly the case for contrast categories that included /r/, such as /r/ vs. /l/, as there are no possible phonemic substitutions with /r/, an alveolar trill, that would meet the criterion of involving a single phonetic feature.

The values in Table 4.3 were calculated as follows. Firstly, for each speaker, the number of occasions on which the contrast error was perceived was divided by the total number of C1 errors for that speaker (and expressed as a percentage). This calculation was performed separately for each of the two error *directions* (e.g., voicing and devoicing). Each directional error percentage was then averaged across the cohort, resulting in an index referred to as the "mean percentage error" (MPE). Finally, the data in Column 2 are the sum of the MPE values for the two directions. For example, the MPE for stop devoicing was 16.1% and the MPE for stop voicing was 2.7%, so the sum was 18.8%. A category was only included if the sum of the two directional MPEs exceeded 2%. Categories that did not meet this threshold, but were reasonably prominent in particular speakers, are discussed in Section 4.4.1. Table 4.4 presents the same type of error analysis for C2.

C1 contrast category	*Mean percentage error*	*Predominant direction (mean percentage error)*
Stop devoicing / voicing	18.8	Stop devoicing (16.1)
Cluster vs. singleton	18.1	Singleton → cluster (11.7)
/l/ vs. /r/	7.6	/r/ → /l/ (4.2)
Glottal vs. null	6.8	/h/ deletion (5.7)
Stop place of articulation	6.3	Stop backing (3.6)
Fricative vs. /r/	5.9	Fricative → /r/ (3.5)
Fricative vs. stop	5.2	Fricative → stop (3.9)
Nasal place of articulation	5.1	Nasal backing (5.0)
/h/ vs. /ɣ/	4.7	/ɣ/ → /h/ (4.4)
Fricative devoicing / voicing	4.5	Fricative voicing (3.4)
Fricative place of articulation	3.0	Fricative backing (2.3)
Nasal vs. stop	2.4	Nasal → stop (1.4)
/v/ vs. /w/	2.3	/v/ → /w/ (1.8)
Consonant vs. null	2.0	Null → consonant (1.3)

Table 4.3. Phonetic-contrast confusions at word-initial (C1) position. The mean percentage error is a measure of the prominence of the error relative to all C1 errors, as described in the text.

As mentioned, the relative prominence of contrast errors in a free-response mode is partly determined by linguistic factors. For example, in the present study, over one-quarter of the target words began with a plosive, so it is not surprising that an error involving this type of phoneme topped the list of C1 confusions (see Table 4.3). However, it can also be seen that stop *voice* errors were considerably more common than stop *place* errors – a result that is unlikely to be an artefact of the word list, as the words beginning with stops were chosen such that in almost every case, they formed a minimal pair based on both voicing and place. However, invoking this type of argument for every contrast category in the table would be a laborious process; therefore, for the time being, the findings in Tables 4.3 and 4.4 will not be interpreted as being informative about speech-production difficulties.

C2 contrast category	*Mean percentage error*	*Predominant direction (mean percentage error)*
Cluster vs. singleton	27.7	Singleton → cluster (22.9)
Nasal place of articulation	27.6	Nasal fronting (19.4)
Consonant vs. null	18.4	Null → consonant (12.9)
Fricative vs. /r/	6.2	/r/ → fricative (5.1)
Fricative vs. stop	6.2	Stop → fricative (5.0)
/l/ vs. /r/	3.5	/l/ → /r/ (3.5)
Nasal vs. stop	2.5	Nasal → stop (2.3)

Table 4.4. Phonetic-contrast confusions at word-final (C2) position.

Tables 4.3 and 4.4 show that it was necessary to define a number of contrast categories that would not have been easily predicted based on the principles of Kent et al.'s (1989) approach: /r/-fricative, /h/ - /ɣ/ and /v/ - /w/. The arguments for and against inclusion of these categories in a Dutch dysarthria assessment are provided in Section 4.4.1. However, for the time being, it is worth making a few observations about how these categories arose. In the case of /r/-fricative, this type of confusion was relatively common, and since it could involve a wide variety of fricatives, often within the same speaker, it was considered justified to group them all as one error / misperception. According to traditional taxonomy, the second confusion, /h/ - /ɣ/, could in fact have been included among the 'fricative place' errors. Indeed, Kent et al. (1989) used a number of words beginning with /h/ as either the target or one of the distractors for this category (e.g., *hill* - *fill*, *feat* - *heat*). However, from a phonetic perspective, this was considered unjustified, as many languages (including English and Dutch) do not involve any constriction within the mouth cavity during the

production of /h/, leading Ladefoged and Maddieson (1996) to suggest that this phoneme should perhaps be regarded as a vowel rather than a fricative. Therefore, it was decided that place confusions involving /h/ would be coded separately in this study (see Section 4.4.1 for a more extensive discussion on this matter). The third substitution that was not predicted in advance is /v/ - /w/. Upon reflection, substitutions of this nature are not surprising, as some accents of Dutch produce /w/ as a labiodental approximant, meaning that the two sounds are more likely to be confused in Dutch than in English. Nevertheless, it is worth mentioning that /v/ - /w/ substitutions may also arise in English, even though Kent et al. (1989) did not test them. Although Johns and Darley (1970) did not observe any /v/-/w/ confusions in their speakers with dysarthria, Platt et al. (1980b) reported four instances of /v/ transcribed as /w/ (where there were 48 productions of /v/ in total, one for each speaker). The direction of this confusion (/v/ → /w/) matches the predominant direction observed in the present study (see Table 4.3).

Figure 4.5 shows contrast-error profiles for four speakers of different intelligibility levels (the same four speakers for whom phonemic errors were displayed in Fig. 4.4). Only errors at the C1 position are shown, due to the fact that milder speakers (in particular) only yield errors for a small number of contrast categories at C2 position. For a given speaker and contrast category, the error value was the number of errors expressed as a percentage of the speaker's total number of C1 errors. The result was only displayed if it exceeded 3%. Thus, categories from Table 4.3 that did not meet this threshold for this group of speakers are not displayed. The categories are grouped according to the phonetic feature that is tested. Thus the first two categories refer to voicing confusions, the next three denote place confusions, the third group involves manner confusions (nasal vs. stop, /r/-fricative and /r/-/l/) and the final group refers to syllable shape. For any given category, the two colours (blue and orange) represent errors in the two possible directions. For example, for voicing categories, the two directions are devoicing and voicing, while for place categories, errors are divided into fronting and backing. If only one colour is displayed, this implies that the perceived errors were unidirectional for the category and speaker in question. As was observed for phonemic errors (Fig. 4.4), the contrast-error profiles are highly individualised. To give a few examples: (1) Speaker 3 is the only speaker for whom a significant proportion of their C1 errors are nasal-stop (N/S) confusions; (2) /r/-fricative (r/F) confusions show a different predominant direction for Speaker 2 compared with Speaker 4; and (3) Speaker 4 shows a different predominant direction for 'stop voicing' confusions than the other three speakers.

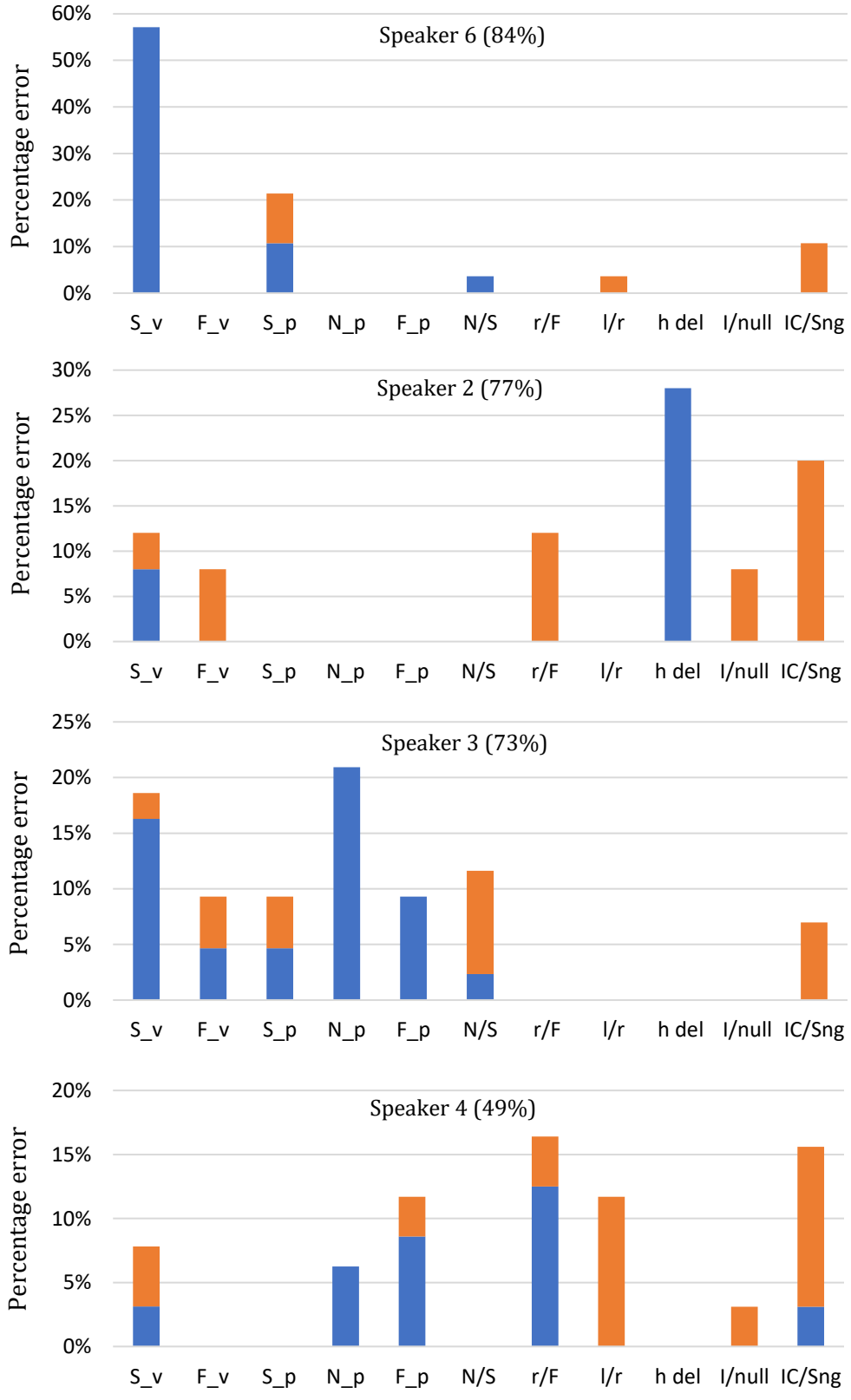


Figure 4.5. C1 error profiles for speakers of different single-word intelligibilities (in brackets). Error rates are calculated as the percentage of the total number of C1 errors for a given speaker. Blue (orange) shading refers to the error direction: devoicing (voicing) for stop and fricative voice errors (S_v and F_v); backing (fronting) for stop, nasal and fricative place errors (S_p, N_p and F_p); nasal → stop (stop → nasal); /r/ → fricative (fricative → /r/); /l/ → /r/ (/r/ → /l/); and deletion (addition) for /h/ deletion (h del), initial vs. null (I/null), and initial cluster vs. singleton (IC/Sng).

4.3.4. Vowel confusions

As argued in Section 4.2, there would be little value in presenting error rates for individual vowel phonemes; therefore this section focuses solely on vowel *confusions*. The original intention had been to condense the inventory of confusions into a smaller number of phonetic-contrast categories, such as vowel duration, front vs. back vowels and high vs. low vowels. However, it was found that most of the common vowel substitutions in the free-response mode did not lend themselves to such simple categorisation, as they involved contrasts in multiple features simultaneously, especially height and advancement. Therefore, categories based on a relatively pure contrast in just one of these dimensions, equivalent to the *feed - food* (front-back) category investigated by Kent et al. (1989), simply did not arise. An exception to this statement is that some of the perceived substitutions that correspond to only a small shift in the vowel space, such as /u/ - /o:/, can be regarded as mainly a vowel-height confusion. However, substitutions between closely-spaced vowels are likely to be of less interest, as they may also arise in the control group. Moreover, the purpose of Kent et al.'s analysis is to reduce *all* vowel confusions to reasonably well-defined phonetic-contrast categories. Since this was not possible, the vowel confusions are presented in terms of the error rates for specific phoneme pairs.

Table 4.5 shows the most common substitutions, where a substitution has only been included if the total MPE, summed over both directions, exceeded 2%. These MPE values were calculated using the same method as that described for consonant contrast-errors. Thus they denote the mean value, across the cohort, of the frequency of occurrence of a given confusion in relation to all other vowel errors for a given speaker. All but one of the substitutions refer to a specific phoneme pair. Substitutions between monophthongs and diphthongs were counted as a single category, as there was considerable variation in the specific phonemes involved in such confusions. It is therefore unsurprising that this category yields the highest MPE. It is also worth noting that the third most prominent confusion, /a:/ - /ɑ/, does, in fact, lend itself to characterisation in terms of a single phonetic contrast. This is because, in the Antwerp accent, these two vowels have almost identical vowel formants (Verhoeven & van Bael, 2002), so they are primarily distinguished on the basis of their *duration*. As can be seen from Table 4.5, vowel shortening was more common than vowel lengthening (by a factor of four).

Vowel confusion	*Mean percentage error*	*Predominant direction (mean percentage error)*
Monophthong - diphthong	24.6	Monophthongisation (17.6)
/ɛ/ - /ɪ/ (*bed - bid*)	10.0	/ɛ/ → /ɪ/ (8.6)
/a:/ - /ɑ/ (*baat - bad*)	8.5	/a:/ → /ɑ/ (6.8)
/ɪ/ - /i/ (*bid - biet*)	7.6	/ɪ/ → /i/ (6.2)
/ɔ/ - /ɑ/ (*bot - bad*)	7.1	/ɔ/ → /ɑ/ (6.2)
/ɑ/ - /ɛ/ (*bad - bed*)	7.1	/ɑ/ → /ɛ/ (4.5)
/u/ - /o:/ (*boet - boot*)	5.8	/u/ → /o:/ (4.3)
/ɔ/ - /o:/ (*bot - boot*)	4.4	/ɔ/ → /o:/ (2.5)
/i/ - /e:/ (*biet - beet*)	3.4	/i/ → /e:/ (2.9)
/e:/ - /ɪ/ (*beet - bid*)	2.9	/e:/ → /ɪ/ (2.4)
/ʏ/ - /ø:/ (*buk - beuk*)	2.6	/ʏ/ → /ø:/ (2.0)

Table 4.5. Vowel substitutions quantified in terms of the mean percentage error (MPE), a measure of the average prominence of the error with respect to all other vowel errors. The final column shows the predominant error direction and the MPE for that direction.

Figure 4.6 shows the substitutions between monophthongs from Table 4.5, but displayed graphically using the F1-F2 vowel space. To produce this graph, the coordinates of each vowel corresponded to the formant frequencies reported for speakers from the Antwerp region in Verhoeven and van Bael (2002). The averages of the values reported for male and female speakers were used. Arrows were then superimposed to depict the most common perceived vowel substitutions, where the thickness of the line is proportional to the mean error rate (shown in Column 2 of Table 4.5) and the arrow head is located so as to indicate the predominant direction. In cases where at least one-third of the errors were in the *non-dominant* direction, two arrow heads are shown.

The final analysis in this section compares vowel error profiles among several speakers. Firstly, inter-speaker variation is of interest in its own right. Secondly, it was possible that the errors of *individual* speakers would show greater consistency (in terms of the directions of the shifts in F1-F2 space) than was observed by averaging across the cohort. For example, one might expect some speakers to show an overall pattern that is indicative of vowel centralisation, in line with acoustic studies of speakers with various types of dysarthria (see, e.g., Kim et al., 2011; Kent et al., 1999; Verkhodanova & Coler, 2018). However, the

degree to which centralisation could, in theory, be observed in the present study is naturally limited by Belgian Dutch phonology. For example, there is only one vowel that is approximately at the centre of the vowel space (*beuk*), and it occurs with relatively low frequency. Therefore, substitutions that include this vowel have a low prior probability of being perceived in a listening paradigm that is confined to real words. Nevertheless, it is still possible that an overall pattern of centralisation could emerge.

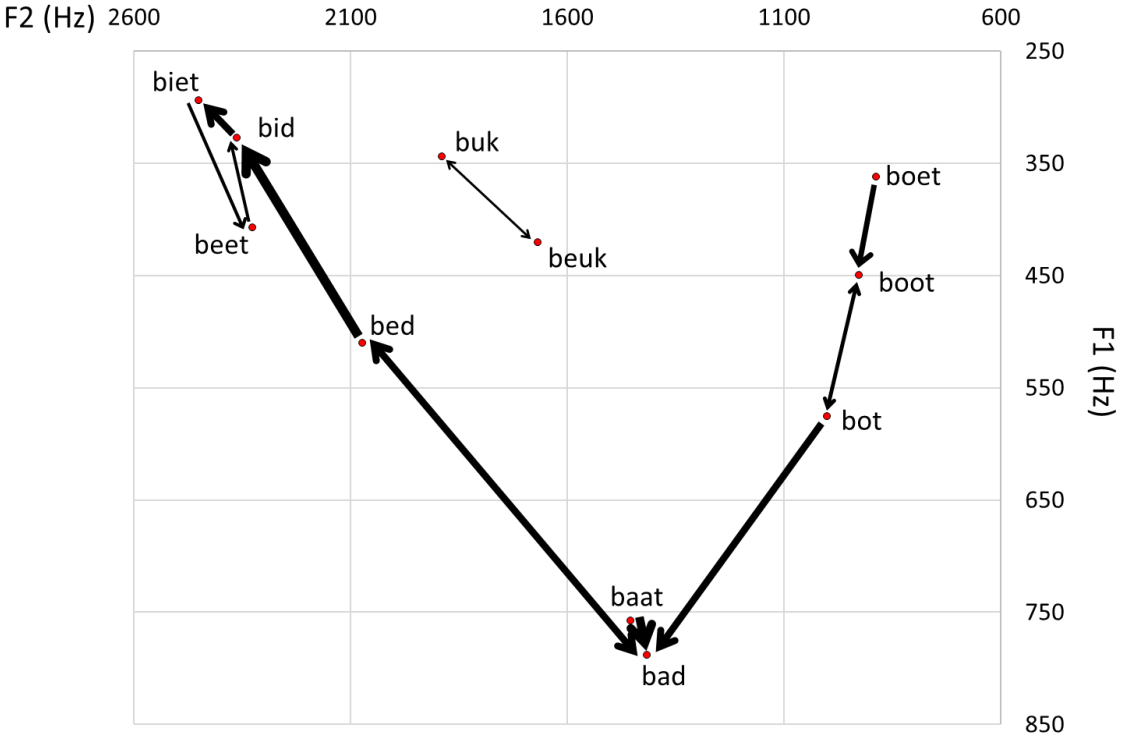


Figure 4.6. Monophthong substitutions portrayed using the F1-F2 vowel space for Antwerp speakers. The thickness of each arrow is proportional to the mean percentage error (summed over both directions), while the arrow head indicates the predominant error direction.

Table 4.6 shows the most common vowel confusions for three speakers (all female). The table includes a confusion between two diphthongs (/ɔu/-/œy/) that was not sufficiently common to appear in Table 4.5, but arose now and again for some individuals. Figure 4.7 shows the same information in graphical form (i.e., in terms of theoretical shifts across the vowel frequency space), for the monophthong substitutions only. On this occasion, the F1 and F2 values for the vowel phonemes were those reported by Verhoeven and van Bael (2002) for the *female* Antwerp speakers.

Firstly, in common with the consonant errors, it can be seen that some contrasts (e.g., monophthong-diphthong, *bed-bid* and *bot-bad*) appear to yield errors in all three speakers, while other confusions (e.g., *bad-bed*, *boet-boot*) are speaker-dependent. Secondly, the individual results largely replicate the averaged results in showing that speakers'

predominant errors are 'diagonal' rather than a pure height or backness contrast. Thirdly, in common with the averaged data in Fig. 4.6, there does not appear to be any overarching trend in terms of vowel height; vowels may be either lowered or raised and the final position may be low, central or high. However, in both the individual and averaged graphs, there is some evidence to suggest that large shifts in F1-F2 space do show a predominant direction in the front-back dimension, namely, vowel fronting.

Vowel confusion	*Total percentage error (error in one direction)*		
	***S7* (92%)**	***S2* (87%)**	***S4* (72%)**
Monophthong - diphthong	20.0 (14.3)	9.8 (4.9)	13.6 (8.4)
/ɛ/ - /ɪ/ (*bed - bid*)	17.1 (17.1)	9.8 (9.8)	9.1 (7.8)
/ɛ/ - /i/ (*bed - biet*)	0	4.9 (4.9)	0
/a:/ - /ɑ/ (*baat - bad*)	14.3 (14.3)	7.3 (7.3)	3.8 (3.2)
/ɔ/ - /ɑ/ (*bot - bad*)	20.0 (20.0)	12.2 (12.2)	14.9 (14.9)
/ɪ/ - /i/ (*bid - biet*)	5.7 (5.7)	19.5 (19.5)	7.7 (4.5)
/ɑ/ - /ɛ/ (*bad - bed*)	0	19.5 (14.6)	14.3 (14.3)
/ɔ/ - /o:/ (*bot - boot*)	2.9 (0)	0	1.9 (0)
/i/ - /e:/ (*biet - beet*)	2.9 (2.9)	0	1.9 (1.9)
/u/ - /o:/ (*boet - boot*)	0	7.3 (0)	9.0 (8.4)
/ʏ/ - /ø:/ (*buk - beuk*)	8.6 (5.7)	0	3.2 (0.6)
/e:/ - /ɪ/ (*beet - bid*)	0	0	1.3 (1.3)
/ɔu/ - /œy/ *(bout - buit*)	0	7.3 (7.3)	1.3 (1.3)

Table 4.6. Vowel confusions for three female speakers with different vowel accuracies (shown in brackets). The first number is the total percentage error summed over both directions. The number in brackets is the percentage error for the direction indicated by the phoneme order in the first column. For example, for the *bed - bid* confusion, the vowel confusions were unidirectional for S7 and S2, while S4 yielded an error rate of 7.8% for *bed* → *bid* and 1.3% for *bid* → *bed*. The top five confusions for each speaker (i.e., summed over both directions) are shaded in grey.

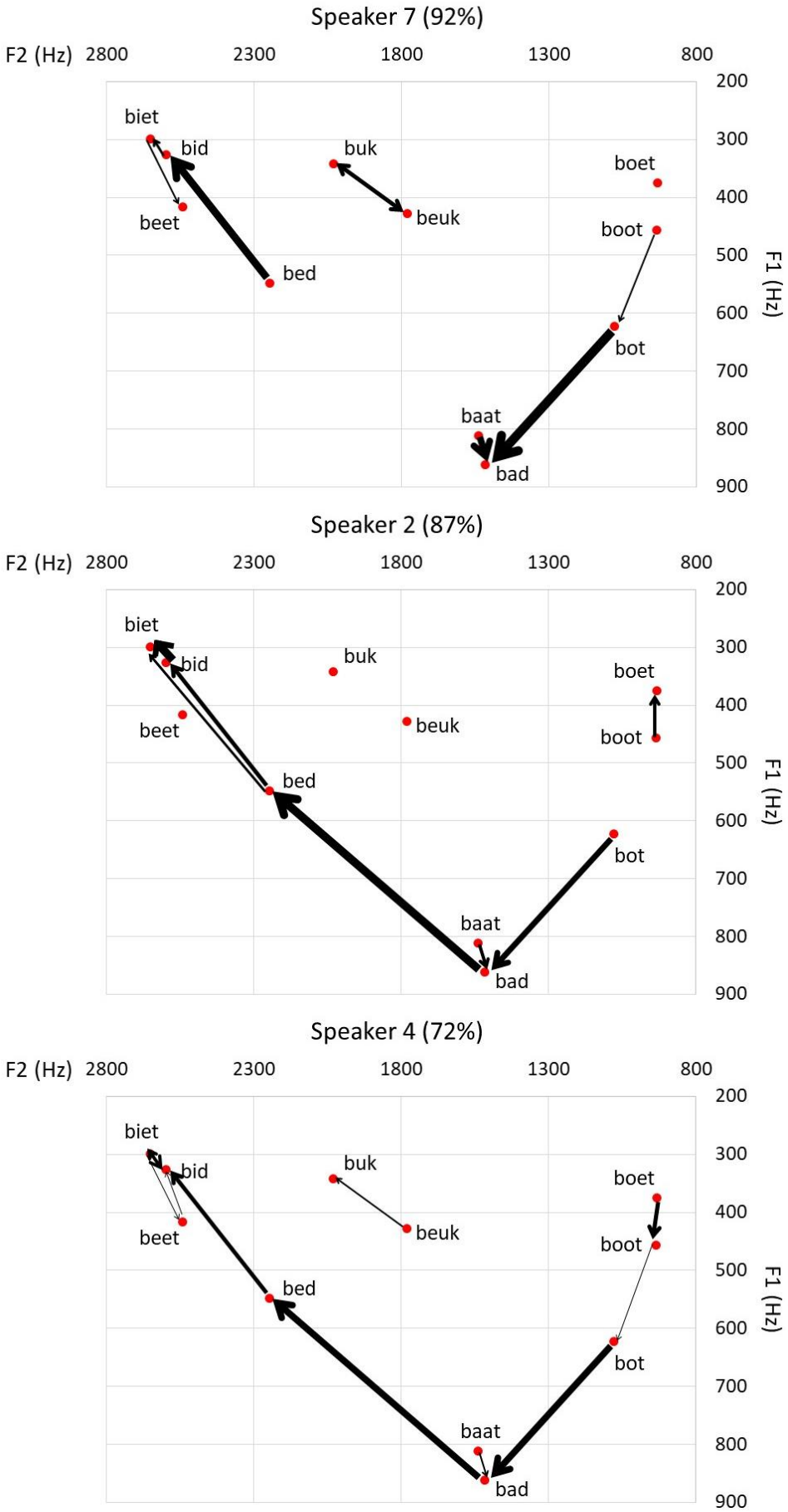


Figure 4.7. Monophthong substitutions for three female speakers.

4.3.5. Inter-listener variability

Before presenting an analysis of the degree of inter-listener agreement for the single-word stimuli, it is worth discussing the meaning of this concept. In an observational assessment, the usual goal is to achieve near-perfect agreement among listeners for the outcome measure of interest, with coefficients of at least 0.8 generally regarded as acceptable. In the present study, there were indeed occasions where the word was perceived identically by all listeners. For example, all five listeners agreed that Speaker 4 produced the word /kɪn/ ("chin") as /kɪnt/ ("child"). However, most of the single-word utterances were perceived in a variety of ways. A brief reflection on the nature of the assessment reveals that this is unsurprising. As documented in Chapter 2, the vast majority of misarticulations produced by speakers with dysarthria are thought to be distortions rather than phonemic substitutions, deletions or additions. It therefore seems likely that listeners might differ in their view of which phoneme is closest to the acoustic signal, especially when additional factors (such as word frequency, which influences the prior expectation of hearing a word) are at play. For example, Speaker 4's realisation of the word /bɛt/ ("bed") was heard by three out of five listeners as /bɪt/ ("pray" or "beg"), while the remaining two listeners transcribed the target. It is possible that the higher lexical frequency of the target may have contributed to the variable response; nevertheless, the fundamental cause of the lack of agreement is almost certainly the fact that the production was a distortion rather than a substitution. Thus the level of inter-listener variability may actually be informative about the nature of an individual's speech deficits, as it could indicate their propensity for distortion errors. Further discussion on this topic is provided in Section 4.4.4, including the clinical implications; however, for the present purposes, the above commentary about the *causes* of listener variability is a useful backdrop for examining the findings.

As argued in Chapter 2 (Section 2.5.1), the level of agreement should ideally be calculated based on the final outcome measure – the profile of phonetic-contrast errors. However, this was not possible, as most listeners did not transcribe the full set of words for a given speaker. In fact, there was considerable variability across the speakers in terms of the total number of independent listeners who transcribed their set of single-word utterances. Therefore, a method needed to be chosen based on the variability in listener responses to individual words. One possibility would have been to use a metric based on whole-word transcription. However, for the more severe speakers in particular, this measure would have been unduly pessimistic, masking any points of agreement among the non-identical transcriptions. Consider an example where the word /pɛn/ was transcribed as /pɪnt/, /bɛnt/, /pɪn/ and /pɛns/. Judged on a whole-word basis, these responses would yield zero agreement, whereas in fact, two of the listeners perceived the same vowel substitution and

three out of four listeners reported a word-final cluster. In the light of this argument, it was decided that the current study would depart from previous research in which listener agreement was calculated at the whole-word level (e.g., Kim et al., 2010). Instead, a novel method of measuring inter-listener agreement was devised. Each word that was transcribed incorrectly, even if this was just by one listener, was subdivided into its three constituent phonemes, and the total number of phonetic-contrast errors[8] among the listeners' responses was calculated. Thus in the example /pɛn/ → /pɪnt, bɛnt, pɪn, pɛns/, there were a total of 6 phonetic-contrast errors (1 on C1, 2 on C2 and 3 on C3). Of these errors, the total number of confusions that were "unique" (only heard by one listener) was established. In this instance, there was just one unique error – the initial voicing of /p/ heard by the second listener. Note that uniqueness was judged with respect to the phonetic-contrast categories defined in this chapter. Therefore, the fact that the fourth listener transcribed a different final consonant from the first two listeners (/s/ instead of /t/) was considered to be unimportant, as in all three cases, the contrast category (final singleton → cluster) was the same. The last step was to calculate the number of *non-unique* errors as a proportion of the total number of errors, which in this case would be 5/6 or 83.3%. Therefore, 83% of the perceived errors for this target were transcribed by at least two independent listeners and can be considered to be somewhat consistent.[9] An overall consistency measure for each speaker was obtained by averaging this metric across all target words (see Table 4.7).

The average consistency across the speakers (± 1 SD) was 61.0 ± 9.5%, meaning that on average, 61% of the contrast errors for a given speaker were heard by at least two listeners. The errors that were "unique" (only reported by one speaker) often arose when the word was perceived as the target by all but one listener, and the final listener transcribed a word that was phonetically similar to the target, e.g., /wɛn/ → /wɪn/. Thus, in these cases, the error would probably be best regarded as (at most) a mild distortion, even though one listener recorded a substitution. Due to this phenomenon, there was a moderate negative correlation, r = -0.60 (two-tailed p = 0.06), between intelligibility and consistency, as speakers of higher intelligibility are, on average, more likely to produce mild distortions. However, there were also other scenarios that resulted in inconsistency, including the

[8] A phonetic-contrast error refers to a confusion that can be defined by a single contrast category. Therefore on the occasions where the substitution consisted of two simultaneous confusions (e.g., /d/ transcribed as /p/, a combination of fronting and devoicing), *two* errors were counted.

[9] The number of listeners who transcribed each word varied between 3 and 5. Thus, the greater the number of listeners, the higher the chance that a given error would be transcribed by at least two listeners. This is one of several reasons as to why the analysis of inter-listener agreement presented in this study should be considered preliminary in nature.

"opposite" situation, i.e., when a speaker misarticulated a word to such an extent that it did not closely resemble any real word of Dutch. Unsurprisingly, this led to considerable variation in listeners' transcriptions. The third situation that encouraged inconsistencies was vowel confusions. Since the vowel errors in this study were not amalgamated into categories (with the exception of monophthongisation and diphthongisation), a distorted vowel that was transcribed differently by different listeners, e.g., /vʏl/ → /ve:l, vul, vœyl/, received a consistency score of zero even though some of the transcriptions might have shared common features (in this case, for example, they all involved lengthening). Thus the consistency score could be considered biased in the sense that speakers who are more prone to vowel errors would be expected to yield a lower score.

ID	*Diagnosis*	*No. of listeners per word*	*Mean word accuracy (%)*	*Mean consistency (%)*
1	ALS	3	78.0	52.3
2	CVA (suspected location: brainstem)	3	77.3	66.7
3	Tumour + surgery (left cerebellum)	3	73.2	69.8
4	Tumour + surgery (fourth ventricle)	5	49.1	68.9
5	Progressive cerebellar atrophy	4	70.6	63.6
6	CVA (right cerebellum)	3	83.9	55.2
7	CVA (pons / left cerebral peduncle)	4	88.4	65.6
8	Cortical watershed CVA (PCA / MCA)	5	68.8	57.8
9	CVA (left cerebellum)	4	62.0	69.4
10	Suspected ALS	3	90.7	40.5

Table 4.7. Mean consistency score (final column) for all speakers. The word accuracy for each speaker is also shown. There is a negative correlation (Pearson's r = -0.60, p = 0.06) between these two quantities such that speakers who were more intelligible yielded errors of lower consistency.

4.3.6. Word frequency analysis

The final analysis presented in this Results section is an examination of word frequency – a potential confounding factor that could affect the relationship between error rate and the phonetic-contrast category. The first step was to rate the lexical frequency of each target on a 5-point scale. The scale was derived based on an online word-frequency database for spoken Dutch (Corpus Gesproken Nederlands, CGN). This database lists the rates of

occurrence of the 5000 most common words in the corpus. Of the 117 target words tested in the present study, 29 did not appear in the top 5000. These words were automatically assigned a frequency rating of 5 (least frequent). The remaining words were categorised into four groups of approximately equal size based on the CGN word-frequency data. The group sizes were not identical because this criterion was considered less important than that of creating groups that were as distinct as possible. This principle is illustrated in Table 4.8, which shows target words that all had similar frequencies, on the border between ratings 3 and 4. In other words, to create two groups of *approximately* equal size, the threshold value for the number of occurrences needed to be placed somewhere in the range 122-186. It was decided that the threshold would be placed at a frequency ≥ 156, as this represented the greatest change between any two consecutive frequency values (see final column). The final group sizes for the five frequency levels ranged from 19 to 29 words.

Word	*English translation*	*Number of occurrences*	*Increase in no. of occurrences*
goud	gold	186	8
Kees	Kees (a name used in a common idiom)	178	0
pil	pill	178	12
maan	moon	166	10
toon	(I) show /exhibit / display	156	0
pen	pen	156	25
rijst	rice	131	7
keus	choice	124	1
bot	bone	123	1
zout	salt	122	11

Table 4.8. Word-frequency data (taken from the CGN) for a subset of the target words used in this study. The bold line indicates the boundary between frequency ratings of 3 and 4. The final column shows the increase in the number of occurrences relative to the next most common target word.

The next step was to calculate the error rate for each of the target words. This was defined as the number of times that the word was transcribed incorrectly divided by the number of times that it was transcribed correctly, calculated across all listener-speaker observations. The mean error rate (± 1 SD) for the set of words in each of the five frequency groups was then calculated and the results are displayed in Fig. 4.8. On average, there appears to be a slight increase in error rate as lexical frequency decreases. However, the correlation between word frequency (on a scale from 1 to 5) and error rate, calculated using the raw data (i.e., the list of 117 words), was weak: Spearman's *rho* = 0.15, *p* (two-tailed) = 0.06.

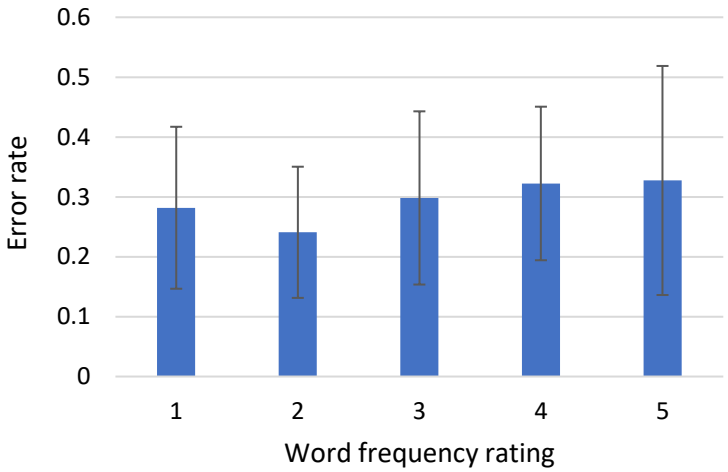


Figure 4.8. Mean error rate (± 1 standard deviation) for the five word-frequency ratings, where a rating of 1 denotes the highest lexical frequency.

4.4. Discussion

4.4.1. Feasibility of phonetic-contrast analysis of consonant substitutions

The first objective of this study was to contribute to the evidence base for Kent et al.'s (1989) method of characterising dysarthric errors, particularly for a language other than English. More specifically, the goal was to determine whether the range of phonemic-substitution errors observed using orthographic transcription could be adequately represented by a reasonable number of phonetic-contrast categories. This subsection discusses the evidence in relation to consonants.

It was stated in Section 4.3.3 that the vast majority of consonant substitutions ($\gtrsim$ 78%) could be captured using a reasonable number of phonetic-contrast categories, such as could form the basis for a Dutch dysarthria assessment equivalent to Kent et al. (1989). The following discussion demonstrates how the final set of consonant categories was chosen. In so doing, it improves understanding of the nature of the information that may be obtained from phonetic-contrast analysis.

The percentage error rates for the most common contrast categories were presented in Tables 4.3 and 4.4 for initial and final consonants respectively. The following discussion (in Table 4.9) reviews the *full* set of contrast categories that were considered for inclusion in a Dutch dysarthria assessment given the findings of the present study. This is a combination of (a) the categories proposed by Kent et al. (1989) that were found to be relevant in the current cohort and (b) additional categories that emerged from the data and were not very low in number and/or confined to just one speaker. The discussion in Table 4.9 focuses on four issues:

(1) *Definition of the contrast category*. Firstly, for some categories, the definition was not as straightforward as the name of the category would imply. In particular, there was sometimes justification for including confusions that comprised contrasts in two phonetic features simultaneously, where the second feature was often voice. This coding strategy was also used by Haley et al. (2000), presumably because voicing confusions frequently occurred alongside manner and place errors. Another scenario in which the category required careful definition was when it pertained to fricatives. The decision in question was whether or not to count /h/ among this consonant class.

(2) *Subcategories*. In some cases, it is discussed whether the category should be divided into subcategories, e.g., based on the direction of the error, or the word segment to which the error applies (C1 or C2). The decision to subdivide was taken when the pattern of errors seemed to be markedly different for each subcategory. This could indicate that a different mechanism is governing error production and/or perception in each subcategory (a *qualitative* difference) or that the mechanism is the same, but there is an important difference in error *frequency*. This is similar to the logic used by Kent et al. (1989), who devoted a category to the 'alveolar-palatal fricative' contrast despite the fact that it could be subsumed by another category – 'fricative place'. Presumably, the authors reasoned that since the alveolar-palatal fricative contrast is known to be particularly vulnerable in dysarthria, it is worthwhile including sufficient test stimuli to measure its error rate with reasonable reliability.

(3) *Error directionality*. Table 4.9 also discusses whether the predominant error *direction* observed for each category (see Tables 4.3 and 4.4) is likely to reflect a genuine difference in directionality (from a speech production or perception perspective) or whether it is likely to have been strongly influenced by functional load. In cases where the directionality is reasoned to be genuine (denoted by bold typeface), it is expected that the same directionality will be observed in the multiple-choice study. Therefore, one can regard the statements containing words in bold in Table 4.9 as hypotheses to be tested in Study 3.

(4) *Phoneme-specific categories*. Kent et al. (1989) included 'categories' that in fact pertained to just two phonemes (e.g., /r/ - /l/). In the present study, even more of these types of confusion were discovered, including some that can be considered 'semi-specific' (e.g., /r/ - fricative). From a practical perspective, it may not be possible to test all such confusions in a dysarthria assessment. Therefore, the following table discusses which of these confusions are most worthy of inclusion and which should be discarded. Future research, with a larger number of subjects, may reach different conclusions.

Contrast category	*Observations*
Voiced-voiceless consonant (word-initial)	This category applies to both plosives and fricatives in Kent et al. (1989). Yet a difference between the two manner classes was observed in the present study: devoicing was more common in plosives, while in fricatives, the converse was observed. However, very few Dutch words begin with devoiced fricatives (which would probably have the effect of making devoicing less likely), while for plosives, there was ample opportunity for errors in both directions; therefore, **devoicing** is likely to be the more common error once functional load has been accounted for. Nevertheless, even if the difference in error directionality between fricatives and plosives is deemed to be artefactual, the voicing contrast is produced in a different way for the two types of consonant, so it could still be justified to define two separate categories. For the present purposes, however, a single category was assumed.
Fricative place	This category applies to both C1 and C2 position. In both positions, backing was more common. The word list provided ample opportunity for the fronting of alveolar and velar fricatives; yet such errors were rarely observed. Therefore it is hypothesised that **backing** is indeed the predominant effect. Due to the scarcity of words beginning with devoiced fricatives, it was decided that minimal pairs involving a *voicing* contrast alongside a place contrast would be included in this category. In particular, words beginning with the devoiced velar fricative are very infrequent, meaning that the backing of word-initial /s/ or /ʃ/ can only be transcribed as voiced. A further point to note is that confusions with /h/ were not coded in this category, as explained in Section 4.3.3.
Stop place --- Nasal place	Kent et al. (1989) employed a single category to examine place accuracy for stops and nasals. In the present study, however, some differences were observed. In C1 position, nasal place confusions only consisted of backing, while stops yielded place errors in *both* directions (albeit more backing). At C2 position, there were almost no 'stop place' errors, while 'nasal place' confusions, especially /ŋ/ → /n/, were common. Due to these differences, the Kent et al. category has been subdivided into 'stop place' and 'nasal place'. As mentioned in Chapter 3 (and discussed further below), the nasal place contrast is thought to be difficult to perceive. Thus the different trends for 'stop place' and 'nasal place' errors could, at least in part, reflect different underlying *causes* of the confusions (production vs. perception). Regarding the error directions, in the case of stops, the predominance of backing may be artefactual; the directionality was not particularly strong and phonology may have played a role (e.g., words beginning with /k/ did

	not always form a minimal pair with /t/). For nasals, the most common occurrence was for the target to be perceived as /n/. This did not appear to be an artefact of the phonology. For example, /n/ → /m/ substitutions were *possible* for every target word beginning with /n/, but never observed. Since /n/ was transcribed for both bilabial and velar targets, no prediction about the directionality of 'nasal place' errors is made.
Stop-fricative	In word-initial position, the stopping of fricatives was the predominant error direction, while the opposite directionality was found for word-final position. However, there were many more opportunities for frication than for stopping in the latter case, as one-third of the word list ended in a stop. Therefore, no prediction is made about the predominant direction. Regarding the definition of this category, note firstly that the labiodental place of articulation only exists for fricatives and not for stops; therefore substitutions between a labiodental fricative and a bilabial stop were considered 'close enough' to constitute a pure manner contrast and to belong to the stop-fricative category. Secondly, note that there were two speakers for whom several /v/ → /p/ substitutions were observed. Although this was recorded as a multiple contrast in the present study (devoicing and stopping), future research may determine that it is justified to include /v/-/p/ minimal pairs as stop-fricative confusions, e.g., if the functional load of /v/ → /b/ is much lower than that of /v/ → /p/.
Stop-nasal	**Denasalisation** was more common than nasalisation at both word positions. Given that stops appeared more often than nasals in the word list, and that words beginning with stops were chosen such that they almost always contrasted with nasals, the directionality of this category is hypothesised to be genuine. The incidence of simultaneous devoicing and denasalisation was high, e.g., /m/ perceived as /p/ and /n/ as /t/. This was partly due to the fact that word-final plosives are routinely devoiced in Dutch. Therefore, simultaneous voice errors were allowed alongside a stop-nasal confusion.
Glottal-null	Although this was a common error in the present study, many of the instances arose from the same speaker. Nevertheless, /h/-deletion has been identified as a hallmark of some types of dysarthria (e.g., ALS). Therefore, it is an important category to include in a Dutch dysarthria assessment (with the exception, perhaps, of accents of Dutch that exhibit frequent h-dropping). The **deletion of /h/** was found to be almost six times as common as /h/ addition. Although there were more opportunities for /h/ deletion, the strength of the directional effect suggests that it is genuine.

Initial consonant-null	This was not a common error (only 3 instances of initial-consonant deletion and 8 instances of **initial-consonant addition** across all speaker-listener observations). There were many more opportunities for initial-consonant deletion in the word list; thus, the directionality of this result is likely to be meaningful.
Final consonant-null	This was a prominent error. **Final-consonant addition** was more common, even though there were many more opportunities for deletion.
Initial cluster-singleton	This was a frequent confusion, with the transcription of a cluster in place of a singleton being the more common finding. However, there were many more opportunities for errors in this direction.
Final cluster-singleton	This was the most prominent C2 confusion. The perception of a cluster instead of a singleton was observed almost five times as often as confusions in the opposite direction. However, this could reflect the fact that singletons comprised 87% of the words that ended in a consonant.
/r/-fricative	This category was defined so as to include substitutions with all fricatives, whether voiced or devoiced. In C1 position, errors occurred in both directions, while in C2 position, /r/→ fricative substitutions dominated. It is not surprising that this contrast shows different behaviour at the two positions; word-final /r/ alters the quality of the preceding vowel, and word-final fricatives are always devoiced. These differences (as well as others) could affect both the likelihood and the directionality of perceived substitutions. Further research is needed to determine whether this category should be subdivided according to word position. For the present purposes, only one category was defined, so as to increase the likelihood that a statistically robust estimate of the error rate would be obtained in the multiple-choice study. Substitutions between /r/ and /h/ were included, despite the argument that /h/ is not a true fricative. It may be more meaningful to consider this category in perceptual terms, e.g., when a speaker's attempt at a trill is distorted, it can perhaps sound like any phoneme that has a noisy production, including /h/.
/r/-/l/	For Belgian-Dutch speakers, /r/ tends to be realised as an alveolar trill, which differs from /l/ in more than one phonetic feature. However, there are no possible phonemic substitutions with /r/ that *would* involve a contrast in a single phonetic feature, and the /r/-/l/ confusion was a prominent error. Thus, it was deemed worthy of inclusion. Most of the /r/-/l/ errors were observed at C1 position, for which there was no strong directionality. As argued for /r/ - fricative, further research might suggest that /r/-/l/ confusions should be subdivided according to word position.

/ɣ/ -/h/	Minimal pairs involving /h/ and other fricatives were not considered to be true place errors. Therefore, /ɣ/ vs. /h/ would need to be a category in its own right. This confusion, which only applies to the C1 position, yielded 17 errors, 14 of which were in the direction /ɣ/ → /h/. Further research is required to determine whether this category should be included in dysarthria assessments, but it was not explicitly tested in the present multiple-choice study. The confusion is discussed further in Section 4.4.3.
/v/ - /w/	Confusions between /v/ and /w/ apply to C1 only. Although the contrast was not considered prominent enough to be explicitly tested in the present study, further research may reveal that it is important. There were 12 /v/-/w/ confusions, 9 of which were in the direction /v/ → /w/.
/j/ - fricative	The most common substitutions for /j/ were fricatives, including /h/. Due to the low frequency of words beginning with /j/, this was not a prominent confusion and it is not included among the final set of contrast categories. However, the error rate for /j/ itself was relatively high: 11.3%, which was higher than the error rate for /r/ in C1 position, 9.8%. Therefore, further research is required to judge the potential importance of /j/, as well as contrasts involving /j/, for dysarthria assessment, since errors involving this phoneme may prove to be of diagnostic value.
/n/ - /l/	This category did not meet the minimum MPE for inclusion in Tables 4.3 and 4.4. However, considering that (a) the category is phoneme-specific and (b) errors were directional (strongly so for the C1 position, where all but one were in the direction /l/ → /n/), the number of errors (20) was not insignificant. Thus, although the category was not explicitly tested in the multiple-choice study, further research may reveal that it is important.

Table 4.9. Consonant contrast categories observed reasonably consistently in the present study. The second column discusses various aspects of each category, as explained in the text. Bold typeface denotes that the predominant error direction observed for the category is hypothesised to be genuine (i.e., not an artefact of the linguistic features of the word list).

Having reviewed the full set of consonant contrasts that were observed with reasonable consistency, it becomes possible to discuss the trade-off between efficiency (i.e., designing an assessment that allows only a manageable number of contrast categories) and comprehensibility (i.e., ensuring that the assessment captures the vast majority of substitution errors perceived in dysarthric speakers). The first thirteen categories listed in Table 4.9 (i.e., every category up to and including /r/-/l/ confusions) were regarded as the minimum set of contrast categories that should be included in future dysarthria assessments. Although some of these categories yielded very low error rates (e.g., nasal vs. plosive), there were often individual speakers for whom the error was more prominent,

and the current population sample is limited both in size and in dysarthria severity and type. Furthermore, the word list has yet to be optimised, so it could turn out that different targets are needed in order to encourage errors in some of the categories. Third of all, a category that is not prominent in an orthographic-transcription study may still produce a high error rate in a multiple-choice paradigm, where functional load is no longer a confounding factor. For all these reasons, the current recommendation is to include at least the first thirteen categories shown in Table 4.9, even those that produced low MPEs. If only these categories are selected, then the proportion of consonant substitutions that cannot be coded,[10] either because they fall outside these categories or because they span more than one category simultaneously, is 15.3%, 14.2% and 22.0% for the three most severe speakers: S4, S9 and S8 respectively. The least severe of these speakers (S8) yielded the highest proportion of non-codable errors. This is because S8 happened to yield a large number of /ɣ/ - /h/ and /v/ - /w/ confusions (12 in total). Allowing these two confusions to be coded would reduce the proportion of uncoded errors to 11.9% for Speaker 8 (the other two speakers did not yield any errors in these two categories). These findings are in broad agreement with those of Haley et al. (2000), albeit for a different population (American English speakers with aphasia and, in some cases, apraxia of speech). These authors showed that only 15% of errors identified by orthographic transcription could not be described by one of Kent et al.'s (1989) vowel and consonant contrast categories (see Section 2.1.5 for more details). Nevertheless, it is important to be aware of the fact that speakers of lower intelligibility than those recruited in the present cohort would be likely to yield a higher proportion of non-codable consonant errors.

Table 4.9 discussed whether some of the categories should be separated into subcategories and whether some of the rarer phonemes of Dutch (e.g., /j/) should be tested on a greater number of occasions than would be warranted based on their natural occurrence in everyday speech, as they may be markers for dysarthria.[11] In the present study, these decisions were often made on practical grounds. However, further research may reveal that there would be clinical value in designating many more categories than that suggested in the present thesis. If so, then a possible solution (to avoid the assessment becoming too time-consuming) would be to administer it in two stages, the first of which would indicate

[10] The errors involving /j/ have been disregarded from this calculation because if the set of contrast categories is not going to allow for the identification of errors involving /j/, then it would not make sense to include target words beginning with /j/ in the assessment.

[11] It is worth noting that the acoustic features of /j/ served as an explanatory variable in van Nuffelen et al.'s (2009b) phonemic model of intelligibility; see Chapter 2, Section 2.2.

the most problematic phonemes or phonetic contrasts, and the second of which would measure error rates for these selected deficits with a greater degree of accuracy.

To conclude this subsection, it is worth summarising some of the perspectives used to formulate the arguments in Table 4.9. Firstly, the analysis emphasised the importance of considering both production and perception arguments when defining phonetic-contrast categories. The Kent et al. (1989) categories were referred to using names that imply a *production* deficit, and the goal of their research was to identify errors that can be related to specific articulatory deficits, such as velopharyngeal insufficiency. The present study, through the use of an empirical approach, demonstrated that a rigid adherence to choosing categories based on production deficits[12] would result in common substitution errors remaining uncoded. The analysis of errors using a free-response mode yielded categories that would not, perhaps, have been predicted based on production arguments (e.g., /r/ - /l/ and /r/ - fricative), but need to be included in a dysarthria assessment since they occur with high frequency. It is likely that some of these categories, such as /r/ - fricative, can be more easily described in terms of an interaction between a production deficit and a perceptual confusion, rather than a pure production deficit (and certainly not one involving a single articulatory gesture). A consideration of the role of production versus perception is also useful when defining categories involving /h/, as this phoneme behaves much like a vowel in terms of its production, but it has *perceptual* characteristics of both vowels and fricatives. A second insight illustrated by the analysis is the importance of taking phonological factors into account when defining contrast categories. For example, it was argued that the 'fricative place' category should permit a simultaneous voice contrast, to allow for the fact that words beginning with devoiced fricatives in Dutch are rare.

4.4.2. Feasibility of phonetic-contrast analysis of vowel substitutions

The findings for vowels were less supportive of Kent et al.'s (1989) methodology. With the exception of durational errors, vowel confusions did not lend themselves to categorisation based on a single phonetic or articulatory dimension (e.g., tongue height, tongue advancement or lip rounding). Rather, the predominant vowel substitutions involved a contrast in two or three of these qualities simultaneously; e.g., *bot* → *bad* (/ɔ/ → /ɑ) is a combination of lowering, fronting and unrounding. It is possible that this finding is specific to the Dutch language. For example, Dutch lacks a vowel similar to English /æ/, meaning that the confusion /æ/ - /ɛ/, which would be a relatively pure height error in many accents

[12] The use of this language is not intended to imply that, in any given instance, a production deficit can be proven to be the cause. Rather, it refers to the definition of errors based on articulatory features such as voice, place and manner.

of American English (see Fig. 4 in Clopper et al., 2005), cannot be observed. The closest approximation in Belgian Dutch, an /ɑ/ - /ɛ/ confusion, involves a large *horizontal* shift in conjunction with the vertical shift (see Fig. 4.6). In the study by Platt et al. (1980b), which investigated speakers with cerebral palsy from the Sydney area, the substitution of /æ/ with /ɛ/ was indeed the second most prominent monophthong confusion. In common with American English, this is predominantly a shift in vowel height in the Sydney accent.[13]

Further research would be required to understand the relationship between the vowel system of a language and the nature of the phonemic substitutions observed via orthographic transcription. Nevertheless, the present findings seem logical based on theoretical arguments: although the cardinal vowel system attempts to describe vowels in terms of the features tongue / jaw height, tongue advancement and lip roundedness, it is known that a change in one of these features rarely occurs without a simultaneous change in one of the others. Therefore, vowel features do not have the same distinctive status as the articulatory dimensions for consonants (voice, place and manner). Furthermore, it is recognised that the vowel quadrilateral is an abstraction and does not represent a direct mapping of tongue position (IPA, 1999, p.12). In their study of speakers with dysarthria due to ALS, Kent et al. (1990) reported very few errors in the front-back vowel category. They remarked that this result was unexpected and needs to be reconciled with the contradictory results of oromotor assessments in this population. However, the vowel pair used to assess tongue advancement in Kent et al.'s test (*feed* - *food*) is a pure error in backness involving a horizontal shift across almost the entire vowel space. The findings of the present study agree with those of Platt et al. (1980b) in showing that most of the perceived vowel substitutions in dysarthria involve confusions between vowels that are reasonably close together in the vowel space for neurotypical speakers. Depending on the vowel system of the language, such substitutions may involve a simultaneous height and backness contrast. Therefore, it seems likely that the finding referred to as "contradictory" by Kent et al. (1990) was an artefact of their multiple-choice distractors, which only allowed for a particular type of advancement error to be observed.

The implication of the present findings is that future Belgian Dutch dysarthria assessments may need to devise an alternative approach for coding vowel confusions – i.e., one that is not based on phonetic features such as height and backness. Further research would be required in order to develop such a method, but one possibility would be to continue to

[13] https://www.mq.edu.au/about/about-the-university/faculties-and-departments/medicine-and-health-sciences/departments-and-centres/department-of-linguistics/our-research/phonetics-and-phonology/speech/phonetics-and-phonology/australian-english-monophthongs

express vowel errors in terms of specific phonemic substitutions (as was done in the present study), but to test only those confusions that (a) are likely to be most important for real-world intelligibility in the sense that they involve common contrasts of Dutch, and (b) are "important" in dysarthria, meaning that the contrast yields appreciable error rates (at least in some dysarthric speakers) and is not observed in neurotypical speakers. The issue of vowel categorisation is discussed further in Chapter 8 (Section 8.2.4).

4.4.3. Articulatory errors in Belgian Dutch dysarthria

The second objective of the free-response study was to contribute knowledge on the segmental, articulatory errors of Belgian Dutch speakers with dysarthria. The main findings in this respect are summarised and discussed in the following paragraphs, with a focus on phonemic errors. An in-depth discussion of *phonetic-contrast* errors is deferred to future chapters, for the reasons given in Section 4.2.

The first set of findings to be discussed is the relative vulnerabilities of the three word segments (C1, V and C2). The highest accuracy was obtained for word-final consonants,[14] which is likely to be due to the limited number of final consonants in the Dutch language. However, for two speakers, the highest phoneme accuracy was *not* observed at C2 position. Inspection of the error profiles of these speakers revealed that they yielded a high incidence of a specific type of C2 error: singleton → cluster. This error is discussed in detail in later chapters; however, for the time being it is worth noting that the extra phoneme usually consisted of a homorganic phoneme of a different manner (e.g., /n/ → /nt/). Such errors should be regarded as distortions rather than intrusions, as they probably arise due to a lack of coordination between the articulators when attempting to pronounce the target phoneme. There was no significant difference in accuracy between C1 and V. This is not surprising given that Belgian Dutch has a rich and relatively crowded vowel system. A high error rate for vowels is also broadly[15] consistent with van Nuffelen et al.'s (2009b) model of intelligibility for Dutch speakers, in which many of the phonemic and phonological explanatory features (which were based on acoustic metrics) pertained to vowels. Note, however, that this model was based on speakers with a variety of accents (including Netherlands Dutch), meaning that the observed relationship between vowel features and intelligibility may not be fully relevant to the current population.

[14] Note, however, that only the difference between C2 and V accuracy was statistically significant.

[15] A direct comparison between van Nuffelen et al. (2009b) and the present study is not possible since (a) van Nuffelen et al. based their explanatory variables on *acoustic* features, and the degree of acoustic distortion for a given phoneme may not be highly correlated with the frequency of perceived substitutions, and (b) van Nuffelen et al.'s features were those that were predictive of overall intelligibility, which does not necessarily imply a high error frequency; see Section 2.3.

Turning our attention to specific phonemes, the labiodental fricatives were among the most vulnerable consonants, with common perceived errors being voicing (of /f/), stopping, and place of articulation confusions. In addition, the voiced labiodental fricative, /v/, was sometimes perceived as the bilabial approximant /w/. The labiodental fricatives were also among the six most vulnerable phonemes in Johns and Darley (1970), who only investigated phonemes in the word-initial position. Further investigations that focus more closely on the labiodental fricatives, especially with cross-linguistic populations, would be required to understand why these phonemes are prone to error. However, some possible explanations are considered here. In the case of /v/, articulatory difficulty could play a role, as this phoneme is believed to have a high level of difficulty (e.g., Stokes & Surendran, 2005; Johns & Darley, 1970). In the case of /f/, the predominant perceived error was voicing, particularly for the word /fɛl/ ('intense'), which was frequently perceived as /vɛl/ ('skin', 'sheet of paper'). Given that there are many more words beginning with the voiced (as opposed to the voiceless) labiodental fricative in Dutch, this finding could reflect linguistic factors. However, possible explanations based on speech production and perception can be imagined. For example, prolongation of the voiceless fricative /f/ requires fine control over the lip muscles, and when this is absent, the duration of the fricative might be reduced. Since voiced fricatives have a shorter duration than their voiceless counterparts, this reduction could cause the phoneme to be perceived as voiced.

A second finding with regard to consonant phonemes was that nasals were frequently transcribed incorrectly, particularly in word-final position, with the most common confusion being a change in the place of articulation. Furthermore, the predominant tendency was for the target to be transcribed as the alveolar (/n/). Fronting of the alveolar nasal (/n/ → /m/) was rarely transcribed despite the fact that a real Dutch word corresponding to this confusion usually existed for words targeting /n/. If these confusions (or at least some of them) were true production errors, then the tendency to default to the alveolar position could be regarded as evidence in support of the schema theory of speech production (Schmidt, 1975), according to which motor programmes for the more common phonemes of a language (in this case, alveolars) are better established and hence less prone to disruption. However, this would not explain why place confusions in general were common for nasals and not for other manners of articulation. In Chapter 5, it is demonstrated that 'nasal place' is also a vulnerable contrast in control speakers, suggesting that an explanation based on perceptual salience is most probable. Indeed, there is evidence to support the vulnerability of 'nasal place' from a perceptual perspective, both in Dutch and in English (see Section 5.4.3). However, it is also worth reviewing the evidence for the vulnerability of 'nasal place' in other studies of dysarthric speech errors. These

studies all used expert transcription (either broad or narrow) carried out by members of the research team, meaning that (a) the intended target was almost certainly known, and (b) the perceived utterance did not need to be a real word of the language.[16] In their study of Cantonese speakers with cerebral palsy,[17] Whitehill and Ciocca (2000a) reported that in word-initial position, nasals were the most *robust* manner class. In Kim et al.'s (2010) study of English speakers with cerebral palsy, nasals were the second most robust manner class (note that the authors did not separate the results according to word position). Platt et al. (1980b) likewise investigated the speech of adults with cerebral palsy (from Sidney, Australia). They found that at word-initial position, both of the nasals were very robust and did not produce any place errors. At word-final position, the nasals were somewhat more vulnerable; nevertheless, the total number of place errors was still relatively small – two place errors for each of the three nasals, where each phoneme was tested on 48 occasions (once in each speaker). Johns and Darley (1970), on the other hand, who only assessed word-initial phonemes, obtained very similar results to the present study for their dysarthric speakers. They showed that while /n/ was a relatively robust phoneme, with just three substitutions out of 130 observations (two of which were the place error /n/ → /m/), place errors for /m/ were relatively frequent, such that /m/ → /n/ was the third most common phonemic substitution in their confusion matrix (yielding 12 errors in total). In summary, previous studies seem to have produced mixed findings. However, as mentioned, all of these studies used expert transcription and it is likely that the transcribers were aware of the targets. This may have reduced the opportunity for the perception of nasal-place confusions in comparison with the current paradigm. Given that Chapter 5 shows that 'nasal place' is a vulnerable contrast in neurotypical speakers at both word positions (C1 and C2), it is most likely that the errors of this type constituted misperceptions rather than misarticulations.

The last consonant phoneme that is worthy of discussion is the rhotic, /r/. This was not a particularly vulnerable phoneme at word-initial position; however at word-final position, it was the most error-prone phoneme after the nasals (see Table 4.2). Given that /r/ is realised as an alveolar trill in the Antwerp accent of Dutch, an allophone that is not particularly common in accents of English, dysarthric productions of this sound have not often been discussed. In principle, one would expect trills to generate a large number of errors, as they are complex articulations that require specific aerodynamic conditions to be met; in the case of apical trills, for example, it is thought that lateral tongue bracing against

[16] The author is not aware of any evidence from orthographic-transcription studies.

[17] Cantonese contrasts the three nasals /m, n, ŋ/ at both initial and final position.

the teeth is required to stabilise the tongue tip during vibration (Howson et al., 2015). In the present study, the most common perceived error in word-initial position was /r/ → /l/, although it is worth noting that the majority of these cases involved substitution of /r/ within a word-initial consonant cluster (e.g., /krɔm/ → /klɔm/). The second most common confusion (albeit attributable to just one participant, S4) was substitution of word-initial /r/ with a fricative, usually the voiced alveolar, /z/. At C2 position, substitution with a fricative (usually /s/) was the only common confusion and was observed in several speakers. Consider firstly the substitution with /l/, which consists of a contrast in two features: absence of the trill and lateral (as opposed to central) release. If the target is produced without a trill, it becomes the voiced alveolar approximant (roughly equivalent to English /r/ in Received Pronunciation - RP). Given that the latter is not part of Dutch phonology, it is perhaps unsurprising that the alveolar lateral approximant would sometimes be perceived in its place.[18] The alveolar and lateral approximants have a high degree of acoustic and perceptual similarity and are known to be difficult to distinguish for native speakers of languages that do not place them in phonological opposition. In addition, evidence from the literature on children's speech shows that substitution of the alveolar trill with /l/ is a common developmental error in languages where the alveolar trill is the only rhotic consonant (see, for example, Tomić & Mildner, 2015). The current finding that /r/ → /l/ substitutions were particularly common in word-initial clusters is also logical, as it is consistent with the observation that, for Dutch speakers, alveolar trills are most likely to occur in absolute word-initial position (due to favourable aerodynamic conditions), whereas the trill feature may not be realised in a consonant *cluster*, especially when it follows coronal and dorsal consonants (Sebregts, 2015). Regarding the substitution of /r/ with a fricative, this might occur when a speaker's attempt at a trill is distorted and sounds more like a fricative. Alternatively, a speaker who is unable to articulate a trill might attempt to approximate it by producing another sustained, periodic sound (frication). The notion that this process might constitute a form of 'weakening' would be consistent with the observations that (a) rhotic trills have the propensity to undergo sound change to fricatives and (b) trills are sometimes produced allophonically as fricatives (Howson et al., 2015). Despite these theoretical arguments showing the logicality of /r/→/l/ and /r/→ fricative substitutions, as well as the findings of the present study, previous studies of the alveolar trill in dysarthria have not necessarily reported the same results. In a narrow-

[18] According to the author's own judgment, in some speakers, it was fairly common for the rhotic to be produced without a trill such that it sounded much like the alveolar approximant in RP English. In most of these cases, the listeners had transcribed the phoneme as the grapheme "r". This is unsurprising, as the demographic characteristics of most of the listeners were such that they would have been highly familiar with English and reasonably proficient speakers of the language.

transcription study of dysarthric speakers of Bengali, Chakraborty (2007) showed that alveolar trills in word-initial clusters were most likely to be *deleted*. In a case study of a Norwegian speaker with severe ataxic dysarthria, Nordli (1996) reported that the rhotic, which is commonly produced as either a flap or a trill in Norwegian, was mostly recorded (using broad transcription) as /l/, but otherwise as /t/, /d/ or /n/. Substitution of the trill with an occlusive was also observed in the present study (although not often) and in a case-study of a Spanish speaker with cerebral palsy (Campoy-Cubillo, 2016).[19] Similar to the explanation suggested above for fricatives, it is possible that speakers may use an occlusive as a means of reinforcing the level of articulation when unable to produce a trill. Vandana and Manjula (2015) used narrow transcription to describe the errors of Malayalam speakers with mild ataxic dysarthria. The most common error observed for rhotics (there is both an alveolar trill and an alveolar flap in Malayalam phonology) was "de-rhotacisation"; in other words, the rhotics were distorted but not substituted.[20] In summary, there is some commonality between these studies, not least the fact that the alveolar trill appears to be a challenging sound for speakers with dysarthria. The precise nature of the misarticulations varies, but they all seem to involve either simplification or an attempt at the trill that is distorted and therefore lacks the trill feature. The present study required transcription of a real word of Dutch. Therefore, it would be expected that different substitutions would be perceived compared to studies that were not constrained in this manner.

The discussion thus far has focused on consonant *phonemes* that were vulnerable in the current cohort. In addition, specific consonant *contrasts* emerged as being prone to error. As mentioned, these errors are likely to be strongly influenced by functional load and perceptual distinctiveness. Thus they are more appropriately discussed in the light of the findings of the normal-control and multiple-choice studies, which shed light on these confounding factors. However, to inform those discussions, it is useful to summarise the results obtained in the present study. In the following summary, a contrast is expressed as directional when the predominant error direction accounted for at least two-thirds of the errors. Otherwise, the error is expressed as bi-directional.

The five most common contrast errors at C1 position were stop devoicing, singleton vs. cluster, /l/ vs. /r/, /h/ deletion and stop place of articulation. The corresponding errors at

[19] The transcription method is described very briefly in this study. It seems that naïve listeners were required to transcribe whole words, but it is not clear whether these had to be real words of Spanish.

[20] In fact, the authors reported that there were no substitutions, omissions or additions for any of the target phonemes, which they attributed to the mild level of dysarthria in their subjects.

C2 position were singleton → cluster, nasal fronting, null → consonant, /r/ → fricative and stop → fricative. Contrast confusions that were not sufficiently common to warrant the designation of a category to be tested in the multiple-choice study (because they pertained to specific phonemes) seemed to involve a simplification or weakening of the target. The most prominent examples were /v/ → /w/, /ɣ/ → /h/ and /b/ → /w/. There were also confusions between /n/ and /l/, which showed strong directionality at C1 position (all but one of the errors being /l/ → /n/). It is likely that this error also represents a simplification, as /l/ is generally considered to be more difficult to articulate than /n/. Yet from a production perspective, a more straightforward simplification, e.g., /l/ → /d/, might have been expected. Therefore, it is possible that the confusion partly arose due to perceptual and/or phonological factors. The fact that /l/ shows a moderate level of vulnerability could be considered consistent with the findings in van Nuffelen et al.'s (2009b) study that the phoneme /l/, and the corresponding phonological feature 'lateral', were important predictors of perceptual intelligibility.

The final set of results to be discussed in this subsection pertains to vowel confusions. The most common substitutions were: (1) monophthongisation, (2) /ɛ/ → /ɪ/, (3) /a:/ → /ɑ/ (vowel shortening), (4) /ɪ/ → /i/, (5) /ɔ/ → /ɑ/ and (6) /ɑ/ - /ɛ/. As can be seen, all but one of these confusions was strongly directional (using the same definition as that given above for consonant contrasts). Most of these substitutions involve only a relatively small shift in F1-F2 space and/or are vowel confusions that are also thought to occur in neurotypical speakers from the Antwerp region (Jo Verhoeven, personal communication). Furthermore, the substitutions involve common phonemes of Dutch, meaning that functional load may have played an important role. Therefore, it would be unwise to read too much into these confusions before acquiring data from neurotypical speakers and from the multiple-choice study. Comparison with previous free-response studies would also be imprudent, as these were carried out for different languages with different vowel systems. Thus, it is difficult to draw any general conclusions about vowel errors in Dutch dysarthria at this stage. However, a summary of the broad findings is as follows: (1) The perceived errors usually involved a simultaneous shift in height and advancement; (2) There is no clear evidence that centralisation was an important process; (3) There was no predominant error direction with respect to changes in height, except perhaps in the top left corner of the vowel space, where an overall *increase* in vowel height could be perceived; (4) There was some evidence that the perceived vowels were *advanced* compared to their targets.

4.4.4. Inter-listener variability

To assess the level of inter-observer agreement, a method was devised that involved calculating, for each erroneous word, the ratio of the number of non-unique phonetic-contrast errors (i.e., errors that were transcribed by at least two listeners) to the total number of errors perceived for that word across all listeners. This metric was then averaged across all erroneous words to produce a measure of "consistency" for the speaker. The mean consistency across the speakers was 61.0% (range 40.5 – 69.8%), meaning that, on average, 39% of a speaker's contrast errors were unique to a single listener. However, this finding is likely to be unduly pessimistic from the point of view of future implementation of the technique for a number of reasons: (1) There was considerable variability among listeners in terms of their skill, experience and listening conditions (especially compared to the variability that would exist among a group of clinical practitioners). (2) It is likely that a consistency measure based on the *outcome measure* (the profile of phonetic-contrast errors) would be higher, as different listeners might yield a similar error rate for a given contrast category, but with the errors distributed differently over the target words relevant to that category. (3) The word list developed in the present study has yet to be optimised; therefore, further research is required to determine whether the stimuli are optimal from the point of view of capturing an individual's phonetic-contrast errors with the greatest possible degree of reliability. (4) Since the vowel confusions could not be reduced to phonetic-contrast *categories*, the consistency of vowel errors was calculated with respect to specific phonemic substitutions. Thus any common phonetic features between the vowels reported by different listeners were not captured. Future research could develop a more sophisticated consistency metric for vowel errors, for example, one that codes transcribed vowels that occupy a similar region of the vowel space as the same substitution (e.g., /ɛ/ → /ɪ/ and / ɛ / → /i/) or one that records confusions as partially consistent if they involve some common phonetic processes (e.g., shortening).

Despite the potential for improvement in inter-listener agreement, it seems unlikely that very high reliability in the outcome measures (the error rates of the contrast categories) will be achievable – at least not for all categories in all speakers. This is largely because the majority of misarticulations produced by speakers with dysarthria are thought to be distortions rather than substitutions. In such cases, listeners are required to use their own subjective judgment in classifying the sound, which will be influenced by factors such as (i) their level of caution (i.e., threshold for detecting a substitution error), (ii) their familiarity with the target word and potential distractors, and (iii) the way in which they perceive the other segments of the target. In a research context, these sources of variability may not present a problem. Given a sufficient number of observers, an *average* error profile could

be determined in which a contrast category with a high error rate would indicate that the speaker had considerable and/or consistent difficulty in producing the contrast in question ('considerable' = a severe distortion that would sound like a substitution to most listeners, 'consistent' = arising on most of the relevant targets). Conversely, a low error rate would imply that the misarticulation was a mild distortion and/or a less consistent occurrence.[21] In a clinical context, however, where there is usually only one listener, poor agreement regarding the speaker's error profile could be problematic, as it would imply that the contrast categories selected for therapy are, to some extent, listener-dependent. Further research of inter-listener reliability (as well as intra-listener reliability, which was not investigated in the present study) is required to understand the implications for clinical practice. For example, it was suggested above that contrasts that are highly distorted in an individual tend to be perceived by all listeners, such as in the case of 'final singleton → cluster' errors for Speaker 4. If this is indeed the case, then the use of a single assessor would still allow for the reliable detection of a speaker's most significant distortions. Otherwise, a possible solution might be to combine the assessment with another investigation (e.g., acoustic analysis) to lend support to the perceptual findings. In fact, since all types of perceptual assessment, whether based on narrow transcription, whole-word transcription, or multiple-choice selection, are inherently subjective, additional information from instrumental analysis is likely to be beneficial in many scenarios.

Upon initial reflection, the inverse relationship between consistency and speaker intelligibility may seem contrary to the findings of van Nuffelen et al. (2008), who showed stronger inter-rater agreement for speakers of higher intelligibility in the NSVO phoneme identification task. However, in their study, the level of agreement was a measure of how many judges identified the same phoneme, irrespective of whether the perceived phoneme was the target or not. This means that *correct* transcriptions contributed to the measure of inter-rater reliability. The current metric, on the other hand, describes the consistency with which *errors* are reported. Therefore, it ignores all target words that were perceived correctly by all listeners, which of course, is a more frequent occurrence in speakers of higher intelligibility.

The lower consistency in speakers of higher intelligibility in the present study probably arose from two mechanisms. Firstly, as mentioned above, speakers with mild dysarthria are likely to yield a higher proportion of distortion (as opposed to substitution) errors, and

[21] For the purposes of this discussion, the error rate is assumed to be solely related to production difficulties; thus, it refers to a situation where the effects of functional load and perceptual similarity can be accounted for.

it seems logical that the former would be perceived with greater variability. Secondly, the speakers with higher intelligibility were assigned fewer listeners (see Section 4.3.1), meaning that there was a higher chance that one of the errors would only be heard by one listener. Nevertheless, it is important to point out that an effect in the opposite direction was also observed in the data. The effect in question was when speakers of low intelligibility misarticulated the target to such an extent that it could not be recognised as any word of Dutch, and thus the transcription involved an element of guesswork. This resulted in highly variable responses, the likes of which were rarely seen in speakers of high intelligibility. An example, observed for S8 (the third least intelligible speaker), was the word /braːt/ transcribed as /paːrt/, /staːt/, /daːt/, /pɔt/ and /bœyt/.

4.4.5. Methodological limitations

To round off this discussion, a brief review of the methodological limitations is presented. Firstly, the study was limited in terms of the size of both populations – the listeners and the speakers. Previous studies that have used whole-word transcription have typically assigned at least ten listeners to each utterance, which, as mentioned in the previous subsection, appears to be necessary to deal with (a) phonemes that are produced as distortions rather than substitutions and (b) words that do not closely resemble any real word of the language. In addition, the listeners varied in terms of their skills, experience, motivation and listening environment.

As far as the speakers are concerned, in addition to the relatively low sample size, there were two speakers who had co-occurring conditions that affected their reading abilities. Speaker 8 had left-sided neglect, as became apparent during the sentence-reading task. However, to the best of the author's judgment, this did not have any effect on his ability to read single words. More problematic was Speaker 9, who had a visual impairment that resulted in a reasonable number of reading errors. An attempt was made to exclude these words from the analysis (see Section 3.1.3), but this procedure may not have been 100% accurate. Therefore the data of S9 were inspected to ensure that his errors did not exert undue influence on the main findings. In general, his errors were similar to those made by other speakers. There were three exceptions: (1) He was particularly prone to initial-/h/ deletion and achieved an accuracy of just 24% for this phoneme. He was largely responsible for the phoneme being the fourth most vulnerable phoneme at C1 position; if his data are omitted, then the mean accuracy for /h/ increases from 87.5% to 94.5%, making it one of the most robust consonant phonemes. However, to the best of the author's judgment, Speaker 9's typical reading errors seemed to involve grapheme *substitution* rather than deletion, so it is unlikely that his failure to produce the glottal fricative was attributable to

his visual difficulties. Furthermore, /h/-dropping is a known phenomenon in some accents of Belgian Dutch (de Louw, 2016). (2) Speaker 9 yielded 10 of the 19 instances of /l/ perceived as /r/. However, this was a genuine speech error, as it was consistently heard during his spontaneous speech. (3) The speaker was also responsible for 8 out of the 13 cases of vowel lengthening. This confusion was also likely to be genuine, as otherwise, the visual error would have corresponded to the perception of an additional grapheme (e.g., *prat* read as *praat*), which was not in keeping with this speaker's typical reading errors.

There were also limitations relating to the word list, as it had not previously been tested in a clinical population and therefore was unlikely to be optimal in terms of attributes such as encouraging contrast errors and maximising inter-listener agreement. Examples of features of a word list that might have a significant confounding effect on the outcome measure (in this case, the profile of phonetic-contrast errors) include the phonetic context and the word frequency. Given all the other criteria that had to be taken into account when developing the word list, phonetic context was not considered, and it cannot be guaranteed that some of the findings were not, at least in part, a reflection of the particular distribution of phonetic contexts employed. Similarly, little attention was paid to word frequency when designing the word list, beyond ensuring that none of the words was of exceptionally low frequency (to the extent that some of the speakers and/or listeners might not have encountered it). The analysis in Section 4.3.6 showed that there was almost no correlation between word frequency and word accuracy. It had been anticipated that the greater the word frequency, the greater the likelihood that the word would be articulated correctly. Furthermore, from a perceptual standpoint, listeners might be more likely to choose a common word over a rare word when pronunciation is distorted. The results did not bear out these expectations. A possible explanation is that the aforementioned positive effects of word frequency were offset by the fact that the more common the word, the more common its constituent phonemes and the greater the number of words with high phonetic similarity. Thus, based on the latter mechanism, words of higher lexical frequency are more likely to encourage errors. Irrespective of the mechanism, it is fortuitous that word frequency does not appear to have a substantial effect on the error rate, as it implies that word lists in intelligibility tests do not have to take this factor into account. Nevertheless, as illustrated in Chapter 5 for the word /ʃɔu/ (meaning '(I) haul'), it would be prudent to avoid words of relatively low frequency, especially if they contain low-frequency phonemes, as these might encourage errors in *all* speakers, including normal controls. Hence they may not be informative about impaired speech production in dysarthria.

4.5. Summary

The main goal of this study was to determine whether the range of phonemic-substitution errors observed in Belgian Dutch speakers with dysarthria by means of orthographic transcription could be adequately represented by a reasonable number of phonetic-contrast categories. The findings revealed that in the case of consonant confusions, phonetic-contrast analysis shows considerable promise: in the current cohort, a minimum of 78% of the contrast confusions observed in each speaker could be coded using 13 phonetic-contrast categories. For vowels, on the other hand, the observed confusions did not lend themselves to categorisation based on features such as height and backness. Thus, further research is required to devise a method of categorising vowel confusions in Belgian Dutch speakers with dysarthria. For the present purposes, vowel error rates are calculated for specific pairs of phonemes. The second contribution of the present study was to identify some of the phonemes (e.g., labiodental fricatives) and phonetic contrasts (e.g., stop voice) that were most vulnerable in Belgian Dutch dysarthria. The discussion of these findings was limited, as a correct interpretation would benefit from information about perceptual distinctiveness (Chapter 5) and functional load (Chapters 5 and 6). Finally, the study obtained preliminary information regarding the level of inter-rater reliability for the identification of phonetic-contrast errors by means of orthographic transcription. On average, 39% of a speaker's contrast errors were unique to a single listener, suggesting that the technique may have low reliability in clinical practice. However, there were several reasons for suggesting that this finding is likely to be unduly pessimistic.

The purpose of the subsequent chapter, which applies phonetic-contrast analysis to neurotypical speakers, is to provide more context for the present findings. In particular, it focuses on two objectives: (1) To determine cutoffs for the diagnosis of dysarthria based on accuracy metrics derived from the proposed single-word reading assessment; and (2) To establish whether any of the phonetic-contrast categories identified in the present study show similar error rates in neurotypical speakers, implying that they should not be included in Belgian Dutch dysarthria assessments.

5. Study 2: Orthographic transcription of single words in control subjects

5.1. Aims and objectives

The aim of this study was to obtain normative data for the phonetic-contrast assessment developed in this thesis. In particular, the study focused on two objectives. The first was to calculate metrics of overall intelligibility for the control speakers. This is useful for determining the potential role of the assessment in dysarthria detection. Previous studies (see Chapter 2, Section 2.1.5) have produced mixed findings on the intelligibility of single-word reading in neurotypical speakers, ranging from (near) perfect word-accuracy to scores as low as 80% in some studies for some speakers. However, in studies where errors (or more likely misperceptions) *are* detected, they often consist of vowel-height or vowel-duration confusions, both of which are likely to occur in the Antwerp accent. This led to the prediction that there would be some overlap in the intelligibility metrics of neurotypical speakers and speakers who had been diagnosed with dysarthria based on having reduced intelligibility in spontaneous speech (see Chapter 3, Section 3.1.1), three of whom yielded word-accuracy scores in excess of 80%. The relevant research question, therefore, was to determine the cutoff score for the diagnosis of dysarthria:

What is the threshold for dysarthria detection in Belgian Dutch speakers from the Antwerp region based on metrics of intelligibility derived from single-word reading?

The second objective of this study was to establish whether any of the contrast categories that yielded high error rates in speakers with dysarthria (as reported in Chapter 4) result in similar error rates in neurotypical speakers. If so, then the implication is that such categories are unlikely to be useful in the clinical assessment of dysarthria. Specifically, the following question was posed:

Do any of the phonetic-contrast categories identified in Study 1 yield errors rates that are not significantly higher in speakers with dysarthria than in neurotypical speakers?

5.2. Method

Control subjects were interviewed in their home or place of work. Of the ten participants interviewed, data from two of them (1 F, 1 M) had to be discarded. In one case (a female participant), the acoustic signal contained sharp noise peaks of high amplitude and low pitch that masked the speech information. Since a lapel microphone was used, it is possible that these were caused by persistent, small movements of the participant, resulting in

contact or friction between the microphone and the participant's clothing. The second dataset that had to be rejected originated from a gentleman who worked as a barrister and was accustomed to public speaking. Preliminary analysis showed that he was 100% intelligible, meaning that no information would have been gained from his data. His manner of speaking was not deemed to be unnatural; rather, he seemed to use an "oratory" style (i.e., clear and formal) in all speaking situations. This finding underlines the importance of recruiting neurotypical participants with a wide range of demographic characteristics.

The first analysis conducted in this study was a calculation of summary measures of single-word reading (SWR) accuracy in neurotypical speakers (Section 5.3.1). These metrics included word accuracy, phoneme accuracy and the accuracies of the three word segments (C1, V and C2). A repeated-measures ANOVA was carried out to determine whether the differences across the three segments were significant. The word-accuracy scores were used to calculate a cutoff for the diagnosis of dysarthria using the proposed SWR assessment. Given the small sample size, an elaborate calculation was not warranted, and the cutoff was determined by subtracting a given number of standard deviations from the mean value (depending on the desired confidence level), having first determined that the data passed the Shapiro-Wilk test for normality.

The second analysis (Section 5.3.2) examined error frequencies across the different consonant phonemes of Dutch and compared the results with those obtained for speakers with dysarthria. As in Chapter 4, the error rate for a given speaker was calculated as the number of occasions on which the phoneme was transcribed incorrectly as a proportion of the number of occasions on which it was uttered. However, due to the low (or zero) error rates for most phonemes in most neurotypical speakers, measures of central tendency were not useful – they were either zero or unduly influenced by comparatively high error rates for one or two outlying speakers. Therefore, a semi-quantitative analysis was conducted in which the consonant phonemes uttered by neurotypical speakers were grouped into discrete levels of vulnerability based on (a) the number of speakers who yielded errors and (b) the typical or maximum error rates observed in these speakers. This allowed for a semi-quantitative comparison of phoneme vulnerability between the two groups.

Section 5.3.3 categorises the consonant substitutions perceived in neurotypical speakers using the same set of phonetic-contrast confusions as defined for speakers with dysarthria. The error rate for each contrast category was calculated using the same metric (the mean percentage error). As a reminder, this involved firstly dividing the number of errors for a given category and speaker by the total number of errors yielded by that speaker at the relevant word position (either C1 or C2). Then, for each contrast category, these normalised

errors were averaged across all speakers to yield the mean percentage error (MPE), a measure of the average relative prominence of the contrast error in question. Error profiles for the two speaker groups (i.e., bar charts of the MPEs of the different consonant contrast categories) were compared by means of visual inspection.

Section 5.3.4 mirrors the analysis presented in Section 5.3.3, but for vowel confusions. As was demonstrated in Chapter 4, vowel errors for speakers with dysarthria did not typically lend themselves to consolidation into a smaller number of phonetic-contrast categories. Therefore, with the exception of monophthong-diphthong confusions, MPEs were calculated for substitutions between two specific vowel phonemes.

The between-group comparisons of contrast errors in Sections 5.3.3 and 5.3.4, which were conducted by visual inspection, do not shed light on the second research question of this study, namely whether there are any consonant or vowel contrast categories that were equally prone to error in speakers with and without dysarthria. This is because the MPE is a measure of the *prominence* of the error relative to all other errors at the same word position, so if a category shows a similar MPE value for the two groups, this does not imply that the control speakers yielded, on average, a similar *number* of errors. In fact, a direct quantitative comparison of error rates for the two groups is difficult with the current data, due to a combination of missing words (for some speakers) and low error rates in the neurotypical group. Nevertheless, Section 5.3.5 presents a preliminary analysis that attempts to overcome these challenges, at least for common neurotypical error categories (i.e., where errors are not confined to just one or two speakers). The precise methodology is intricate and is described in Section 5.3.5. However, it essentially involved calculating approximate *vulnerability* rates (i.e., the number of errors divided by the number of occurrences of the contrast) for the common contrast categories, and then implementing either t-tests or Mann-Whitney U-tests (both one-sided) to identify categories for which the vulnerability rate was significantly higher in speakers with dysarthria than in neurotypical controls. A one-sided test was warranted because (a) it was expected that error rates for all categories would be higher for speakers with dysarthria (random sampling noise excepted) and (b) there was no interest in distinguishing between the case of equal vulnerability rates and the case of higher vulnerability in neurotypical speakers.

Having identified a set of contrast confusions that is unlikely to be dysarthric, the final analysis (Section 5.3.6) involved recalculating word-accuracy values for speakers with and without dysarthria while neglecting these errors. That is, words that only contained "non-dysarthric" errors were scored as correct. The updated word-accuracy scores of the control group were then used to calculate new cutoffs for the diagnosis of dysarthria.

5.3. Results

5.3.1. Word accuracy and segmental accuracies

Each word uttered by each neurotypical speaker was assessed by three listeners. Most of the listeners assessed one-quarter of a speaker's utterances. Thus, in total, the data for a given speaker typically arose from the transcriptions of 12 independent listeners. Table 5.1 shows the age and gender of each control subject, along with their scores for word accuracy, phoneme accuracy, and segmental accuracies, calculated in the same way as for the speakers with dysarthria (see Chapter 4, Section 4.3.1).

ID	*Age*	*M/F*	*Word accuracy (%)*	*Phoneme accuracy (%)*	*C1 accuracy: (%)*	*V accuracy: (%)*	*C2 accuracy: (%)*
S11	63	F	86.9	94.8	92.6	95.2	97.0
S12	66	F	93.7	97.4	97.9	97.2	97.0
S13	78	M	85.2	94.4	91.8	93.4	97.9
S14	56	F	95.1	98.3	98.5	97.7	98.8
S15	64	M	94.3	98.0	96.7	97.7	99.7
S16	83	M	92.8	97.2	99.3	94.5	97.6
S17	77	F	89.8	96.4	95.8	95.5	97.7
S18	76	M	87.4	94.9	93.9	96.3	94.7
Mean across cohort ± 1 SD			**90.7 ± 3.8**	**96.4 ± 1.5**	**95.8 ± 2.8**	**95.9 ± 1.6**	**97.6 ± 1.5**

Table 5.1. Demographic information and accuracy metrics for the control subjects assessed using orthographic transcription of the single-word reading stimuli.

It can be seen that there was no difference in mean accuracy between C1 and vowels. The highest segmental accuracy was for word-final consonants, as was also the case for speakers with dysarthria. However, a repeated-measures ANOVA showed that the differences in accuracy across the three segments failed to reach significance: $F(2,14) = 2.75$, $p = 0.098$. In addition to the small sample size, this could be due to the fact that the segmental accuracies are close to the ceiling value. To facilitate comparison of the results in Table 5.1 with the corresponding data for speakers with dysarthria, the word accuracies of all speakers are plotted as a histogram (see Fig. 5.1a), in order of increasing accuracy. The dotted pattern indicates a control speaker, while the filled grey bars represent

speakers with dysarthria. Fig. 5.1b shows the same type of chart, but for C1 (rather than word) accuracy. C1 accuracy was chosen in preference to the other segmental accuracies (V and C2) because it was found to have the highest correlation with spontaneous-speech intelligibility in speakers with dysarthria (see Chapter 7, Section 7.3.1).

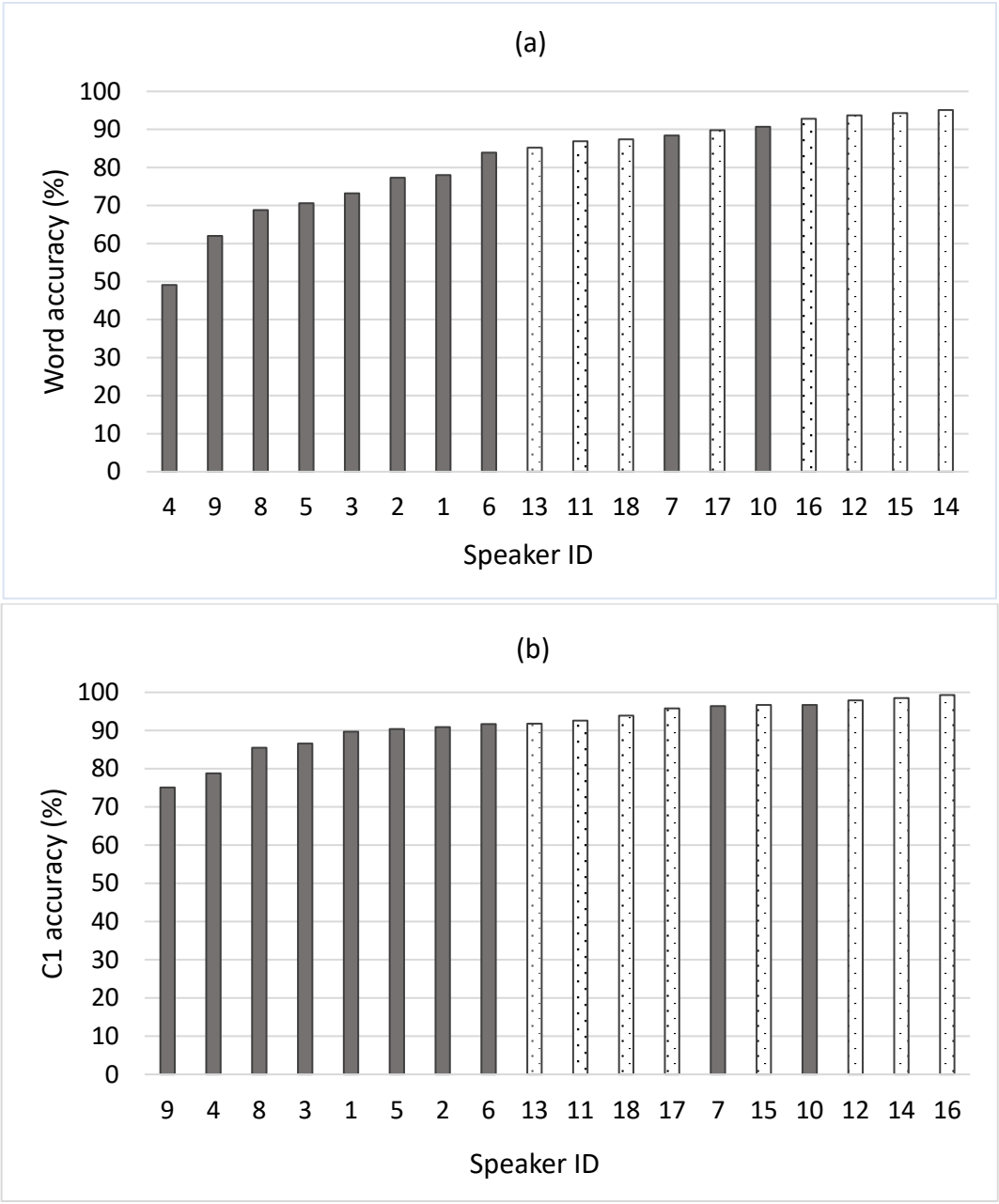


Figure 5.1. (a) Word accuracy and (b) C1 accuracy for all speakers. The dotted pattern represents neurotypical speakers, while the filled (grey) bars represent speakers with dysarthria.

As expected (see Section 5.1), there was overlap in the single-word reading intelligibility metrics of neurotypical controls and speakers who had been diagnosed with dysarthria based on subjective assessment of their spontaneous-speech intelligibility (SSI). Word accuracy is the best choice of intelligibility metric for computing cutoffs because it spans

the largest range of values. Based on the assumption of a normal distribution, 97.5% of neurotypical speakers would achieve a word accuracy of at least 83.2% (1.96 standard deviations below the mean). Such a threshold would result in a diagnosis of "no dysarthria" in 30% of speakers who had been diagnosed with dysarthria based on their SSI. This proportion does not change when the cutoff is based on 1.645 SDs below the mean (95% of neurotypical speakers), which results in a threshold of 84.4%.

5.3.2. Vulnerabilities of consonant phonemes

As mentioned, in neurotypical speakers, low numbers of errors made it difficult to calculate meaningful measures of central tendency for consonant phonemes. For example, all but one of the control speakers yielded perfect accuracy for the word-initial voiced velar fricative /ɣ/, while the remaining speaker (S13) yielded frequent substitutions for this phoneme such that his percentage error was 73%. Simple averaging across the eight speakers would result in a mean percentage error of 9.1%, which would make /ɣ/ the third most vulnerable word-initial consonant. However, since the errors are solely due to one speaker, this would be misleading. The median and modal values, both of which are 0%, would be more representative. However, these metrics are not particularly informative – they completely ignore any outlying data, and in fact, the majority of consonants yielded median and modal values of 0%, meaning that it would not be possible to draw any conclusions about which consonants are most prone to error. In view of these limitations, the vulnerability of each consonant phoneme was conveyed by providing a semi-quantitative description of the error distribution – one that mentions both the *number of speakers* who yielded errors for the phoneme and the typical *error frequencies* observed for those speakers. These descriptions are provided in Tables 5.2 and 5.3 for initial and final consonants, respectively. To aid interpretation of the data, phonemes that show similar error distributions based on the combination of these two metrics have been grouped together in one row. Phonemes that were judged to be more vulnerable are situated closer to the top of the table.

The percentage error values cited in the tables, since they refer to individuals rather than to the whole cohort, should be interpreted with caution. This is particularly the case for phonemes that were only tested using a small number of targets. For example, consider a phoneme that was tested using three targets. A percentage error of 33% would mean that a total of three errors were perceived out of nine tokens (since there were three listeners per word). As can be seen from the tables, such an error rate would be considered comparatively high. Yet it could arise if the speaker had made a slip-of-the-tongue on just one word, meaning that one mistake results in a comparatively high (and yet "spurious") error rate. Therefore, to give the reader an impression of the robustness of the percentage

error values quoted for individual speakers, the number of occasions on which each phoneme was tested is shown in parentheses in the first column (this information has been omitted for the phonemes in the final row, as no error rates are reported). The number of instances of each consonant phoneme was calculated using singletons only; words beginning or ending with a consonant cluster contributed to the total number of clusters.

C1 phoneme (# instances)	*Description of the distribution of percentage errors across speakers*
clusters (17)	Yielded at least one error in *every speaker*. The maximum error rate across the cohort was 12%, observed in two different speakers.
/p/ (5) /r/ (8) /t/ (4) /m/ (6) /n/ (3)	Yielded errors in *at least half the cohort*, but generally only with very low frequency. The highest error rate seen in any speaker ranged from 17% (for /r/ and /m/) to 33% (for /n/).
/f/ (2) /v/ (5) /l/ (6)	Each of these phonemes yielded errors in *3 out of 8 speakers*. The maximum error rates for the fricatives were 33% for /f/ and 13% for /v/ (both observed in the same speaker). For /l/, all three speakers who yielded errors produced an error rate of 11%.
/ɣ/ (5) /h/ (9)	Generally perfect or near-perfect accuracy. However, *one speaker* (S13) yielded a high error rate (73%) for /ɣ/ as well as a moderate error rate (33%) for /h/.
/b/ (10) /d/ (8)	Notable errors were observed in only *one speaker* (S11). Her error rate was 13% for both /b/ and /d/.
/j, k, s, ʃ, w, z/	Perfect or near-perfect accuracy in all speakers

Table 5.2. Semi-quantitative descriptions of the distributions of C1 error rates across the cohort of neurotypical speakers. The data are organised such that phonemes closer to the top of the table are more vulnerable, to the best of the author's judgment.

The findings in Tables 5.2 and 5.3 should be interpreted with caution and a lengthy discussion of their implications would not be justified. Nevertheless, it is worthwhile reporting these preliminary data, as they can be used for comparison with future studies. Further, it is worth briefly discussing how the more robust findings differ from the findings for speakers with dysarthria. To facilitate this discussion, the phonemic error rates for speakers with dysarthria are repeated here in a condensed form (see Table 5.4).

C2 phoneme (# instances)	*Distribution of percentage errors across speakers*
/m/ (3) /n/ (10) /ŋ/ (3)	Yielded a few errors in *at least half the cohort* (and in the case of /n/, in 7 out of 8 speakers). The maximum error rates across the cohort were 22% for /m/ (obtained in two speakers), 17% for /n/ (obtained in one speaker) and 33% for /ŋ/ (obtained in two speakers).
/l/ (11) clusters (14)	Yielded one or two errors in half the cohort, i.e., *four speakers*. However, the maximum error rates were relatively low: 9% for /l/ and 7% for clusters.
/r/ (13)	Generally perfect or near-perfect accuracy. However, *one speaker* (S18) produced a moderate error rate of 31%.
/p, t, k, f, x, s/	Perfect or near-perfect accuracy in all speakers

Table 5.3. Semi-quantitative descriptions of the distributions of C2 error rates across the cohort of neurotypical speakers. Phonemes closer to the top of the table were judged to be more vulnerable.

Word-initial consonant	*Error rate (%)*	*Word-final consonant*	*Error rate (%)*
/f/	26	/ŋ/	39
/ɣ/	23	/m/	21
/v/	21	/n/	10
clusters	16	/r/, /f/	8
/m/	15	clusters, /l/	7
/h/	13	/s/, /t/	5
/l/, /d/	12	/x/, /k/	3
/b/, /t/, /j/	11	/p/	<1
/r /, /s/	10		
/ʃ/	9		
/w/, /p/	6		
/n/	5		
/k/, /z/	<2.5		

Table 5.4. Mean consonant error-rates for speakers with dysarthria, displayed in order of decreasing frequency. Error rates were derived from Table 4.2, but have been rounded up or down to the nearest 1%. In addition, errors below a certain frequency have been grouped together.

There are many points of agreement between the two groups of speaker. For example, for word-initial consonants: (i) the alveolar fricatives are more robust than their labiodental counterparts; (ii) clusters and the bilabial nasal yield relatively high error rates; (iii) the phonemes /r/ and /t/ have an intermediate level of vulnerability; and (iv) /k/ and /w/ are

among the most stable phonemes. In word-final position, most of the phonemes show very similar rankings in the two groups: the nasals yield the highest confusion rates, while the consonants /p, t, k, s, x/ are very stable. The only clear difference is that the phonemes /r/ and /f/ seem to be more problematic (relative to other phonemes) for speakers with dysarthria than for neurotypical speakers. Note, however, that one of the control speakers yielded a comparatively high error rate of 31% for word-final /r/.

The *discrepancies* between the two groups of speaker lie mainly in the word-initial consonants. Firstly, it can be seen that word-initial /n/ is relatively prone to error (or misperception) in neurotypical speakers, but relatively robust in speakers with dysarthria. In both populations, the perceived errors on nasals were almost entirely due to place confusions (/m/ ↔ /n/). Yet there was only one instance of /n/ → /m/ at C1 position in the dysarthric dataset (n = 10), compared to 7 for neurotypical speakers (n = 8). In the latter case, the errors were distributed over three different speakers and over all 3 target words that began with /n/, suggesting that the errors are not spurious. The reason for the greater number of perceived errors for word-initial /n/ in neurotypical speakers is unknown. Secondly, it can be seen that while confusions involving initial /p/ were relatively common in neurotypical speakers (compared to error rates for other C1 phonemes), /p/ was one of the most stable word-initial phonemes in speakers with dysarthria. As it happens, word-initial /p/ was one of the few phonemes for which an *absolute* error rate could be calculated in neurotypical speakers (i.e., the measure of central tendency was meaningful). The mean error rate across all speakers (7.6%) was, in fact, also higher than the corresponding value for speakers with dysarthria (5.7%). Further investigation showed that in control speakers, all but one of the /p/ confusions (11 out of 12) consisted of /p/ transcribed as /b/. In speakers with dysarthria, on the other hand, the voicing of voiceless consonants was not particularly common. The between-group difference with regard to stop voicing is analysed further in Section 5.3.5. The third notable difference between the two populations concerns the voiced plosives, /b/ and /d/, which were among the more vulnerable C1 phonemes in speakers with dysarthria, but highly robust in the control group. In speakers with dysarthria, the perceived error for these phonemes was usually devoicing – a confusion that rarely arose in control speakers (the total number of 'stop devoicing' errors in the cohort was six,[1] which is low given the very high occurrence of voiced initial plosives in the word list). The final notable difference between the two groups involves the phonemes /ɣ/ and /h/. Both of these consonants were perceived with perfect (or near-perfect) accuracy

[1] This only includes errors on /b, d/ when they appeared as *singletons*. The perception of the devoiced counterpart was more common when /b/ and /d/ were part of a *cluster*.

in all but one neurotypical speaker. However, the remaining speaker, S13, yielded large numbers of errors: 11 out of 15 for /ɣ/ and 9 out of 27 for /h/. In speakers with dysarthria, these phonemes typically yielded lower error rates, but the errors were spread across a number of speakers. Regarding the *nature* of the errors, in the case of /h/, the main process in both groups was deletion. For the voiced velar fricative, a variety of confusions was observed. In the case of the neurotypical speaker, /ɣ/ was transcribed as /k/ on 7 out of 11 occasions; otherwise, it was either replaced by /h/ or deleted. In speakers with dysarthria, the two most commonly-transcribed substitutions for /ɣ/ were /h/ and /r/, while the remaining confusions consisted of the perception of the cluster /ɣr/. The /ɣ/ → /k/ substitution perceived for the neurotypical speaker was not transcribed at all for speakers with dysarthria. The implications of these findings are discussed in Section 5.4.3.

5.3.3. Consonant contrast errors

In this section, consonant confusions are categorised using phonetic-contrast analysis and quantified using the MPE metric, which represents the *prominence* of the error with respect to all other errors. As argued throughout the thesis, an error rate that reflects *vulnerability* cannot be calculated for a free-response study, as information about the number of opportunities for making each contrast error is too difficult to obtain. One might presume that this problem is solved by acquiring normative data. That is, if two groups (with and without dysarthria) utter the same word list, then the number of potential errors is the same. Thus, it should be possible to make a direct comparison between the two populations based on the number of errors of a specific type. Unfortunately, the situation was not so straightforward. In the case of neurotypical speakers, for most of the contrast categories, the total number of errors across all speaker-listener pairings was very small. As a result, this number was prone to heavy influence by a single, outlying speaker. For example, there were a total of 6 'fricative → plosive' errors in the neurotypical speakers and a total of 7 'nasal backing' errors. However, while all instances of fricative stopping arose from just one speaker (the speaker for whom /ɣ/ was often transcribed as /k/; see Section 5.3.2), the nasal backing errors were distributed across four different speakers. Thus it is clear that, in general, nasal backing is a more likely confusion than fricative stopping, despite the fact that the total number of errors is approximately the same. The *median* number of errors would also be unsuitable as a metric, as this was zero in many cases. Therefore, the only meaningful way of comparing the two groups was using the MPE metric. Recall that to obtain this metric, the normalisation relative to the total number of C1 or C2 errors is carried out *prior* to calculating the group mean. This was found to be necessary in order to reduce the influence of outlying speakers. However, their influence could not be eliminated

completely; thus in the remainder of this section, a finding is only mentioned if, to the best of the author's judgment, it appears to be valid. The MPE does not eliminate confounding factors and it is likely that the highest values will pertain to contrasts that involve phonemes of high frequency, as was the case for speakers with dysarthria. Nevertheless, the confounding factors ought to have a similar influence on all speakers; thus errors that are common in speakers with dysarthria, but not in the control group, can be considered "dysarthric". Figure 5.2 presents the C1 results for both groups. Contrast categories have only been included when the sum of the MPE values in the two directions exceeds 2%.

When interpreting Fig 5.2, it would not be justified to conclude, for example, that neurotypical speakers yielded more 'stop backing' errors ('S_p', blue bars) than speakers with dysarthria. The dysarthric speakers yielded a total of 511 C1 errors relative to 145 in the control group. Thus the number of C1 errors per speaker is approximately 3.5 times higher for the dysarthric group, meaning that, to a rough approximation, the MPE would need to be at least 3.5 times higher in neurotypical speakers than in speakers with dysarthria before one could conclude that the two groups make a similar number of errors. Therefore, the finding that an MPE value is somewhat higher in neurotypical speakers is not particularly interesting. In contrast, when the MPE is higher in *dysarthric* speakers, this indicates that the error is, at least in part, a result of impaired speech production. To summarise, the appropriate ways of interpreting Fig. 5.2 are: (a) comparing *relative* errors (e.g., the ratio of 'stop voice' errors to 'stop place' errors) in speakers with dysarthria vs. neurotypical speakers and (b) highlighting categories for which the MPE is notably[2] higher in speakers with dysarthria. The findings in the following paragraphs are based on these two principles.

Similarities between the profiles for the two groups include the high rankings (i.e., *relative* MPE values) of stop devoicing, initial singleton → cluster, /h/ deletion,[3] and nasal backing; all of these confusions appear within the top five directional errors for the respective populations. In addition, 'stop voice' errors (S_v) are much more common than 'stop place' errors (S_p) in both groups despite the fact that there was an approximately equal opportunity for each type of error to arise.

[2] The word "notably" has been used to emphasise that it would be unwise to draw conclusions from MPE values that are only slightly higher in dysarthric speakers, especially for contrasts that have low MPE values in both groups or were only tested on a few occasions. Therefore, caution was exercised in applying this principle and a finding is only brought to the attention of the reader if the raw data reveal that it was based on a reasonable number of observations.

[3] Note, however, that the high prominence of /h/ deletion in neurotypical speakers was mainly due to a single speaker, who yielded this error on a highly consistent basis.

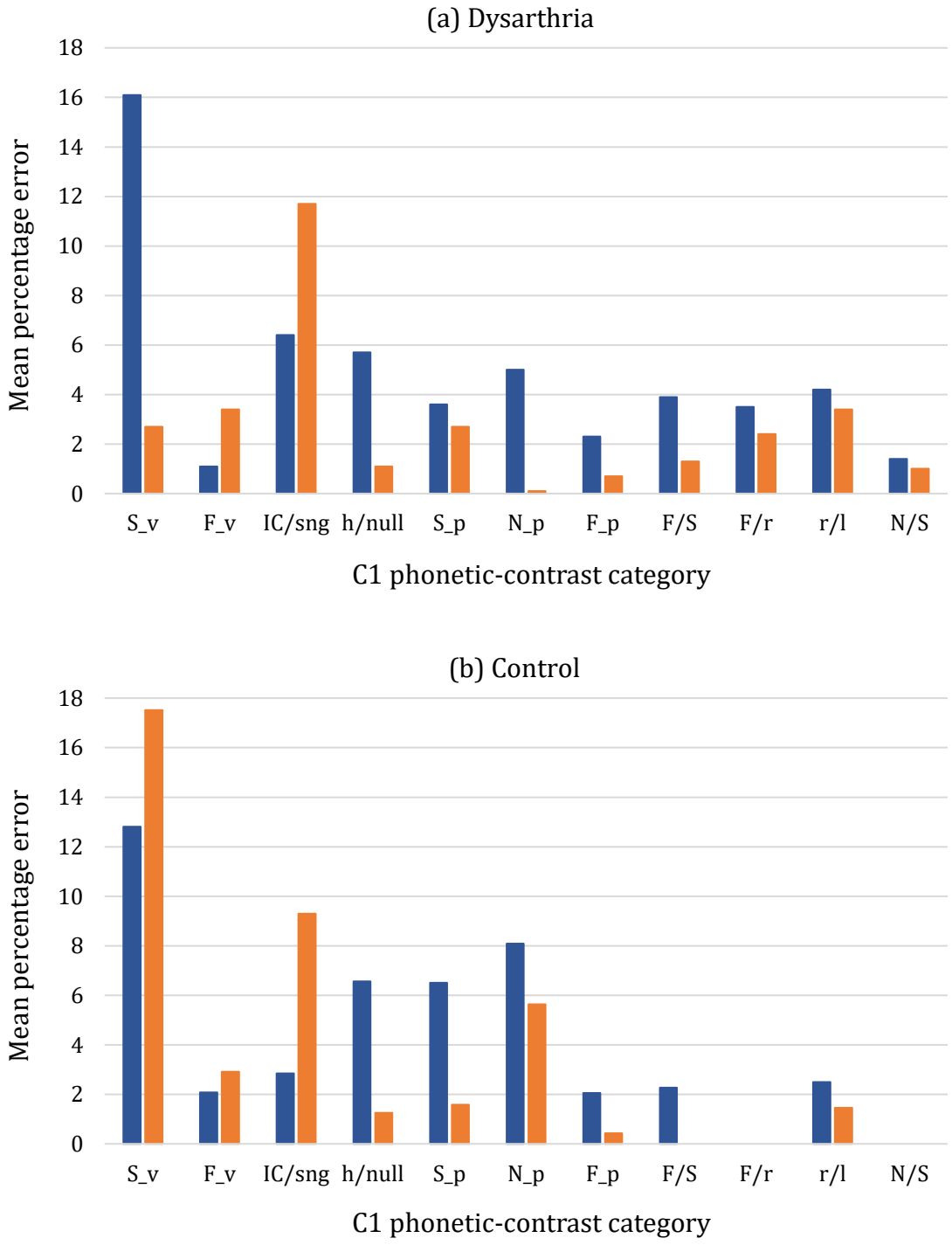


Figure 5.2. C1 contrast-error rates in (a) speakers with dysarthria and (b) control speakers. The two colours denote the two directional errors. Error rates represent the number of times that the directional error was observed divided by the total number of C1 errors made by the speaker. These values were then averaged over the whole cohort to yield the mean percentage error. Blue (orange) refers to the following directions: devoicing (voicing) for stop and fricative voicing errors (S_v and F_v); deletion (addition) for initial cluster vs. singleton (IC/sng) and /h/ vs. null (h/null) confusions; backing (fronting) for stop, nasal and fricative place errors (S_p, N_p and F_p); fricative → stop (stop → fricative) for the category F/S; fricative → /r/ (/r/→ fricative) for the category F/r; /r/ → /l/ (/l/ → /r/) for the category r/l; and nasal → stop (stop → nasal) for the category N/S.

The most notable differences between the two profiles are as follows. Firstly, speakers with dysarthria yielded relatively few *voicing* errors (orange bar) in the 'stop voice' category, while neurotypical speakers were more likely to yield a voicing error than a devoicing error. Since voiced plosives appeared twice as often as devoiced plosives in the word list, and since devoicing is thought to be more natural than voicing, the finding for the neurotypical group is more surprising. It is discussed further in Section 5.4.3. The reader may also notice that the MPE for stop voicing is more than six times higher in neurotypical speakers than in speakers with dysarthria. Therefore, as explained above, it is likely that neurotypical speakers actually yielded more errors of this type. This issue is investigated statistically in Section 5.3.5. A further difference between the two populations, as also mentioned in the previous subsection, is that speakers with dysarthria only yielded nasal backing errors, while neurotypical speakers also gave rise to a reasonable number of instances of /n/ transcribed as /m/ (fronting). Finally, there were two categories seen in dysarthria that did not yield any errors in neurotypical speakers: fricative vs. /r/ and stop vs. nasal. Therefore, these error categories can be regarded as having a dysarthric component. The same conclusion can be reached for stop devoicing, initial cluster → singleton and fricative stopping, on the basis that the MPE values are all notably higher (see Footnote 2) for speakers with dysarthria than for neurotypical speakers.

Figure 5.3 shows the relative prominence of the C2 contrast errors within each population. The total number of C2 errors was 336 in the dysarthric group and 89 in the control group, meaning that, to a very rough approximation, an MPE value would need to be at least 3.8 times higher in the neurotypical group for two groups to yield a similar *number* of errors. For the control group, 'fricative vs. stop' errors did not arise at all. Therefore this error can be considered dysarthric. Substitutions of /l/ with /r/ were also confined to the dysarthric group; however, the MPE is low and most of the errors arose due to one speaker. Apart from these two differences, there is relatively close agreement between the two populations in terms of the relative prominence of the directional errors. Regarding the MPE values themselves, there are two contrasts for which the MPE is notably higher in speakers with dysarthria: 'final singleton → cluster' and 'final cluster → singleton'. This implies that these errors have a dysarthric component.

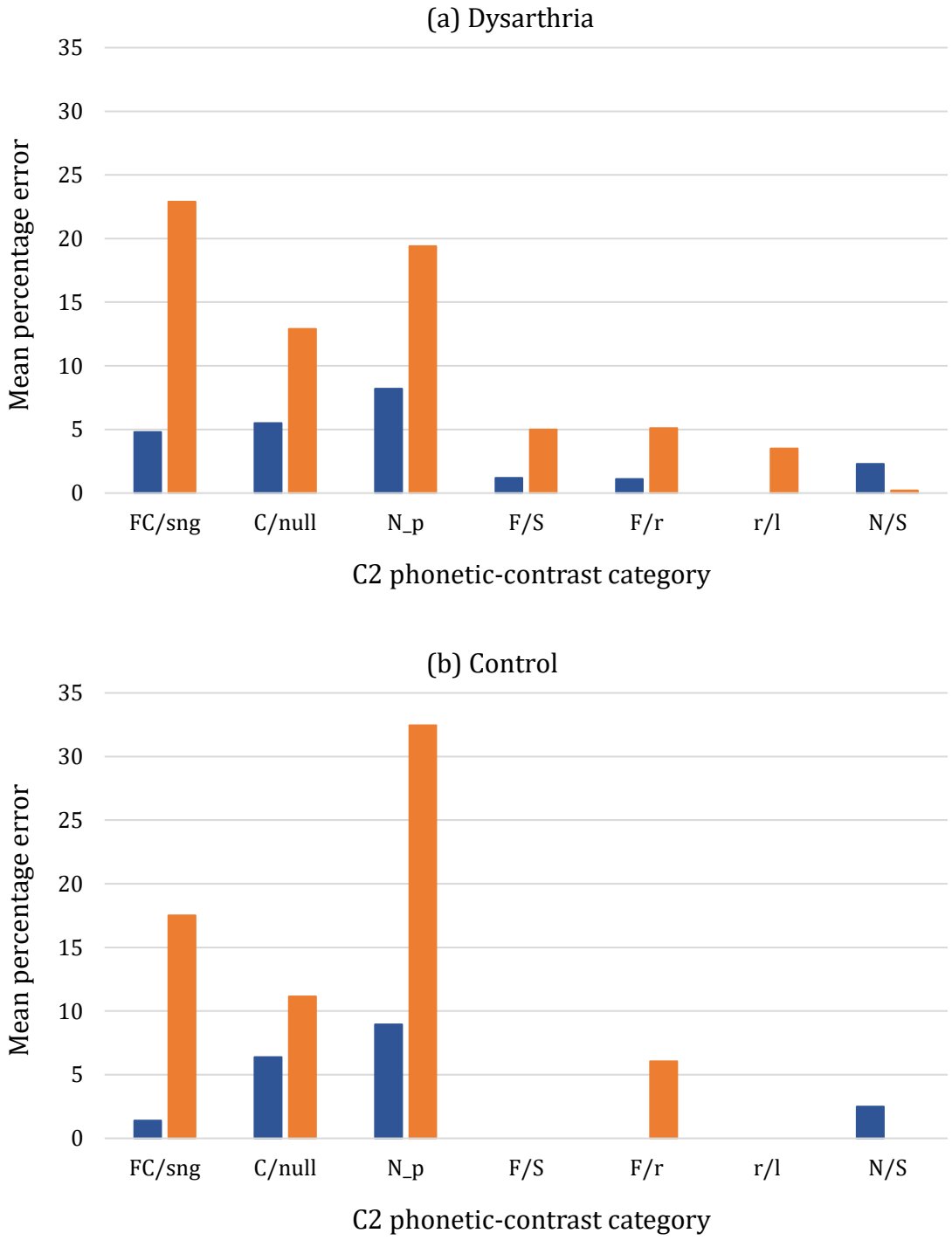


Figure 5.3. C2 contrast-error rates in (a) dysarthric and (b) control speakers. Blue (orange) refers to: deletion (addition) for final cluster vs. singleton (FC/sng) and final consonant vs. null (C/null); backing (fronting) for nasal place errors (N_p); fricative → stop (stop → fricative) for the category F/S; fricative → /r/ (/r/→ fricative) for the category F/r; /r/ → /l/ (/l/ → /r/) for the category r/l; and nasal → stop (stop → nasal) for the category N/S.

Finally, it is interesting to note that for the syllable-shape categories (final cluster vs. singleton and final consonant vs. null), the neurotypical subjects show the same directionality as speakers with dysarthria; i.e., the perception of an intrusion is more

common than the perception of a deletion. For the 'final cluster vs. singleton' category, this is unsurprising, as there are many more opportunities for cluster *formation* in the word list than cluster *reduction*. However, the result for the 'final consonant vs. null' category cannot be explained by functional load. There are far more opportunities for the perception of final-consonant deletion than final-consonant addition; yet the latter is more common. The fact that this is also the case for neurotypical speakers may at first seem important. However, examination of these confusions in the neurotypical population showed that 4 out of the 5 instances consisted of the same error: /ʃɔu/ meaning '*(I) haul*' perceived as /ʃa:l/ meaning '*scarf*'. The perception of this word substitution was very common among all speakers (with and without dysarthria) and is likely to be due to the much higher lexical frequency of /ʃa:l/. Furthermore, it is doubtful that /l/-addition in word-final position signifies a similar production or perception deficit as the addition of other final consonants. Full vocalisation of word-final /l/ is common in different accents / languages of the world. It is also observed as a developmental error in children and as a phonological change over time. Indeed, Dutch has many words in which /l/ has been replaced by the diphthong /ɔu/, as can be seen by comparison with the same words in English (e.g., /zɔut/ - *salt*). Therefore, transcription of the word /ʃa:l/ does not necessarily imply that the token /l/ was perceived, as listeners are accustomed to word-final /l/ being imperceptible in this context. Thus, the instances of 'null → final consonant' observed in neurotypical speakers do not seem to have the same origins or implications as the corresponding errors in dysarthric speakers.

5.3.4. Vowel confusions

Mean percentage errors for vowels were calculated in the same way as for consonants, and are displayed in Fig. 5.4 along with the data for speakers with dysarthria. The total number of vowel errors was 580 in the dysarthric group and 135 in the control group, meaning that, to a very rough approximation, an MPE value would need to be at least 4.3 times higher in the neurotypical group for the number of errors in each group to be equal. It can be seen that there are three vowel confusions that yielded an appreciable error rate in speakers with dysarthria, but not in control speakers: /ɑ/ - /ɛ/, /ɔ/ - /o:/ and /ʏ/ - /ø:/. In particular, the contrast between /ɑ/ and /ɛ/ is the 5th most common vowel error in the dysarthric group (see Table 4.5), and since it involves relatively common phonemes of Dutch, the finding that it was not observed in control speakers is likely to be reliable. Thus the error can be considered dysarthric.

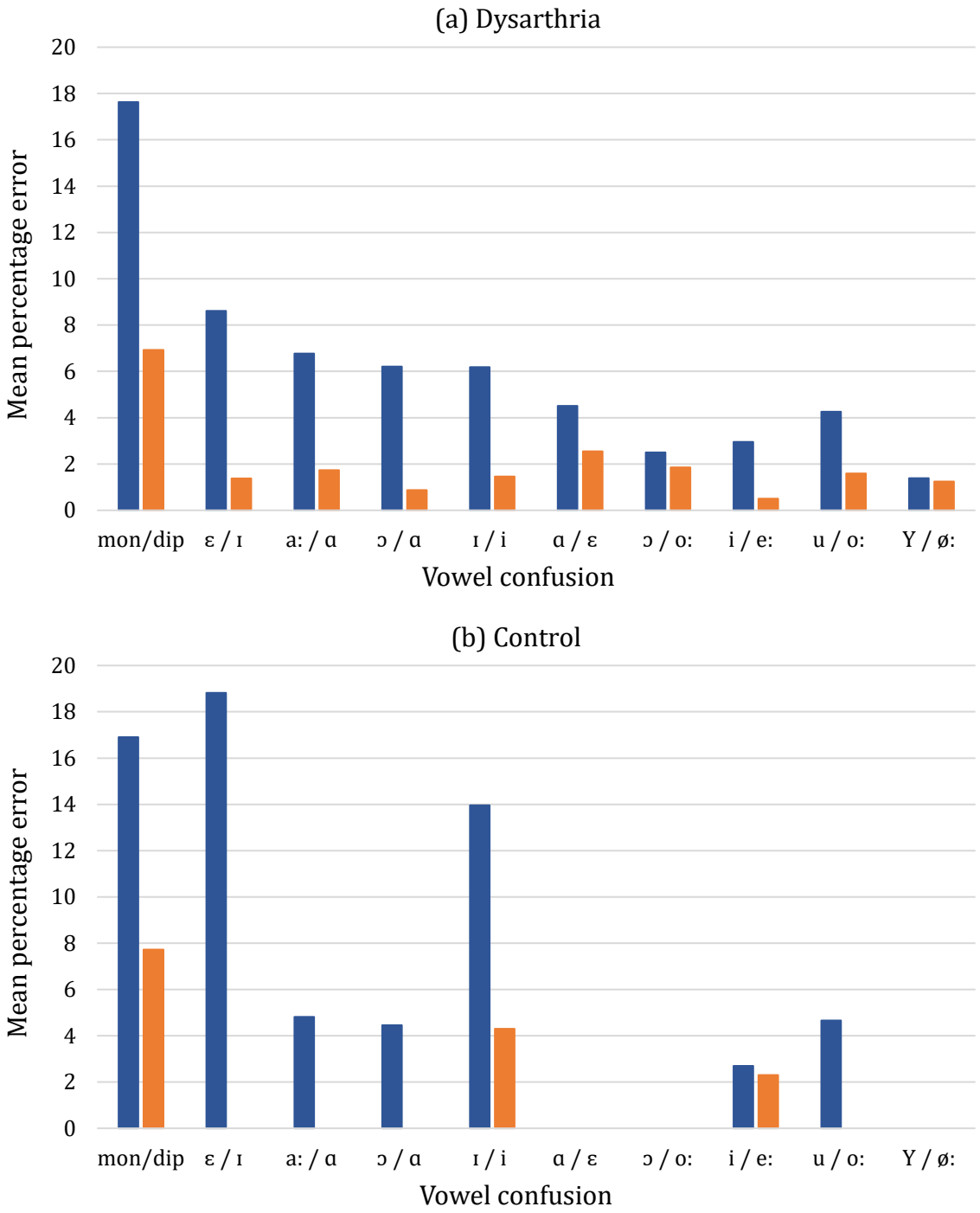


Figure 5.4. Vowel confusions in (a) dysarthric and (b) control speakers. Blue denotes the error direction implied by reading each label from left to right. Thus, the errors are monophthong → diphthong (diphthong → monophthong), /ɛ/ → /ɪ/ (/ɪ/ →/ɛ/), /a:/ → /ɑ/ (/ɑ/ → /a:/), and so on.

There was one confusion (not shown in Fig. 5.4) that arose more often in the *control* population, namely the perception of /ɛi/ as /œy/ (both diphthongs). The MPE for this error placed it lower in ranking than all the other vowel confusions in Fig. 5.4b. However, the total number of /ɛi/ → /œy/ confusions in the control population (8) was not insignificant, considering the fact that /œy/ is not a common phoneme. Furthermore, as

mentioned, this total was higher than the number of errors seen in speakers with dysarthria (5). The diphthongs /ɛi/ and /œy/ trace fairly similar paths across the vowel frequency space, at least in Standard Belgian Dutch (Verhoeven, 2005), and although there is also a roundedness contrast (/ɛi/ being unrounded and /œy/ rounded), some Belgian speakers have no lip-rounding in the first element of /œy/ (Collins & Mees, 2003: p.136). In other words, the confusion is a logical one, involving diphthongs with fairly similar features. It is not known why the error was less common in the dysarthric group. However, the data show that /ɛi/ was frequently perceived as a monophthong in speakers with dysarthria.[4] Thus, the fact that other confusions were perceived for this phoneme (which may have reflected genuine distortions rather than misperceptions) may have reduced the likelihood that it would be perceived as the closely related diphthong.

Thus far, the summary of the findings has focused on vowel errors that were very obviously more prominent for one of the speaker groups. To assess the remaining vowel confusions, the reader is reminded of the general principles for interpreting the data. Errors that have a high ranking within the *neurotypical* population probably arise due to factors such as high functional load and close perceptual similarity between the phonemes. A higher MPE in the neurotypical group than the dysarthric group is uninteresting (unless it is of the order of five times higher), as it does not imply that the control speakers yielded more errors. In contrast, vowel confusions that have a notably higher error rate in speakers with dysarthria are a consequence, at least in part, of disordered speech production. Prominent normative errors include monophthong vs. diphthong, /ɛ/ → /ɪ/ and /ɪ/ → /i/. In none of these cases, however, is the MPE value at least 4.3 times higher than in the dysarthric group. Errors that appear to be "dysarthric" include the /ɑ/ - /ɛ/ confusion mentioned above, as well as /a:/ - /ɑ/ (duration errors) and /ɔ/ - /ɑ/. Finally, Fig. 5.4 shows that 4 out of the 7 vowel confusions that arise in both populations are unidirectional in neurotypical speakers but bidirectional (albeit with strong directionality) in speakers with dysarthria. This suggests that in some cases, it is the atypical *directionality* of a confusion that marks it as dysarthric.

Figure 5.5 redisplays the monophthong confusions as vectors within the Antwerp F1-F2 vowel space, as described in Chapter 4 (Section 4.3.4). Comparison of the graphs for the two populations shows that the dysarthric errors are more varied and include confusions that involve large movements across the vowel space, in particular, /ɑ/ → /ɛ/. The vowel confusions in the control group mainly consist of the perceived raising of vowels in the crowded top-left corner of the vowel space. These confusions could be indicative of

[4] Substitutions of /ɛi/ with a wide range of monophthongs were observed, including /ɛ/, /ɑ/, /a:/, /e:/ and /ʏ/.

phonological changes in progress, as discussed in Section 5.4.3. There is also some evidence of the lowering of back vowels in neurotypical speakers, but this is relatively weak.

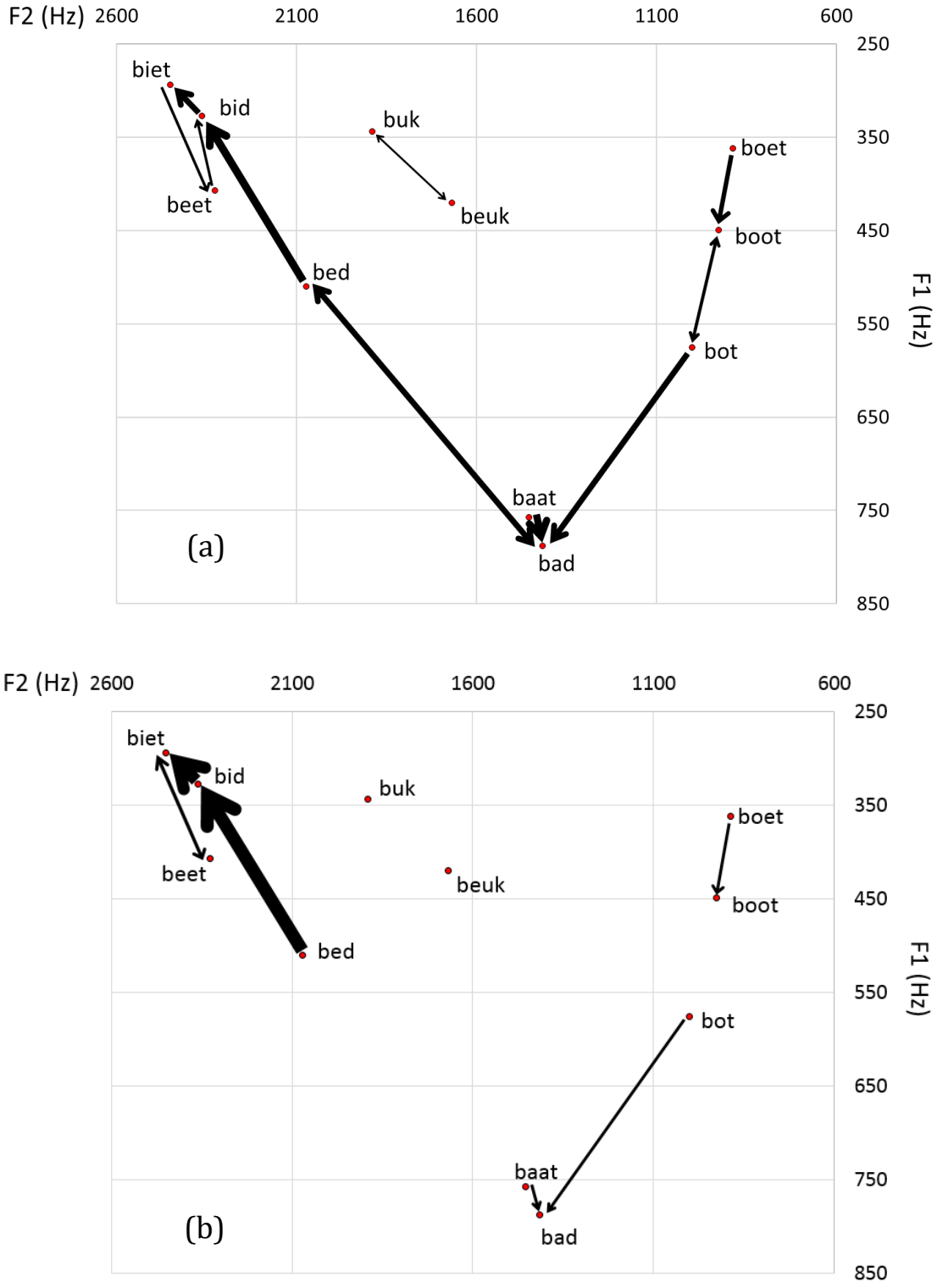


Figure 5.5. Monophthong confusions for (a) dysarthric and (b) control speakers. The thickness of each arrow is proportional to the sum of the MPEs across both directions. The arrow head indicates the predominant error direction; two arrow heads are shown when the errors are bidirectional.

5.3.5. Quantitative between-group comparison of common contrast errors

This section describes a quantitative comparison between the vulnerabilities of common contrast categories in speakers with and without dysarthria, where the term "vulnerability" refers to the ratio of the number of observed errors to the number of potential errors. Since the number of potential errors is difficult to determine in a free-response study, one of the purposes of acquiring the control data was to bypass the need for this information; in other words, if two sets of speakers utter the *same word list*, then they can be compared purely on the basis of the number of observed errors. However, there were two problems with implementing this strategy:

(1) In the case of neurotypical speakers, the total number of errors across all speaker-listener pairings was often small and prone to undue influence by a single, outlying speaker. For this reason, the current analysis is restricted to the most *common* error categories – i.e., those for which errors were distributed over a larger number of speakers.

(2) Most of the speakers in the dysarthric population were not assessed using the full word list (see Chapter 4, Table 4.1). Therefore, the assumption that the two sets of speakers uttered the same word list is not valid. The solution was to obtain an approximate estimate of the number of words uttered by each speaker that tested the contrast in question. For some contrasts, this calculation may appear to be relatively straightforward. For example, 'nasal backing' in C1 position only applies to words beginning with /m/, meaning that it would simply be a case of adding up the number of words beginning with /m/ uttered by each speaker. However, there is an extra layer of complexity, as one of the six words beginning with /m/ ('maan' - *moon*) does not form a minimal pair with /n/. Therefore it needed to be decided whether to exclude such words, on the grounds that 'nasal backing' was not possible, or to include them, since speakers sometimes yielded a contrast error even when no minimal pair existed (e.g., /ma:n/ transcribed as /na:m/, meaning *name*). A further level of complexity lies in the fact that some words were particularly prone to error, while others tended to be realised with high accuracy. Thus all eligible words cannot be considered equal in terms of the opportunity they provide for yielding the error in question. In addition to these challenges that arise for categories such as 'C1 nasal backing', which only applies to target words beginning with /m/, for other contrasts, it would be extremely laborious to identify the exact number of opportunities for yielding an error. For example, in the full word list, there were 95 words that began with a consonant singleton. Thus in order to determine the number of opportunities for 'initial singleton → cluster' errors, one would need to count the number of initial singleton-cluster minimal pairs that are possible for each of these 95 words. In the light of these complexities, and given the preliminary nature of this analysis, it was decided that a relatively simple approach would be taken to

calculating the number of *possible* contrast errors (i.e., the denominator). Firstly, all words containing the relevant phoneme(s) were included, irrespective of whether they formed a minimal pair based on the contrast in question. In addition to the fact that this strategy made the calculation considerably less laborious, it was considered justified for the reason given above (/ma:n/ transcribed as /na:m/). Thus, for word-initial stop devoicing, for example, all words beginning with /b/ or /d/, whether they occurred as singletons or part of a cluster, were included in the denominator. Secondly, the propensity of the word to produce errors was simply ignored. Consequently, if a given speaker had a recording failure for a word that tended to yield errors in most other speakers, the speaker's vulnerability rate would have probably been an underestimation. Conversely, if the omitted word was generally robust across all speakers, then the vulnerability rate was likely to have been an overestimation. Correction for this factor would have been extremely laborious and was considered beyond the scope of the analysis. Finally, in the case of 'singleton → cluster' confusions (both C1 and C2), *all* words contributed to the denominator if they began with a phoneme that was capable of forming consonant clusters according to Dutch phonology. The fact that these words varied in terms of whether they actually produced such minimal pairs (and if so, how many) was neglected.

Vulnerability rates were calculated as follows: the total number of errors observed for a given speaker and contrast category was firstly divided by the denominator (calculated as described in the previous paragraph) and then by the number of listeners. The mean and median vulnerability rates across each cohort (dysarthric and control) were calculated (see Table 5.5). Inspection of the mean and median values, as well as a comparison between them, provides some insight into the error distributions. For example, a median of zero implies that the majority of the cohort did *not* yield the contrast error in question. Finally, statistical tests were performed to examine the significance of the difference between the two groups for each contrast. If the data were found to be normal (using the Shapiro-Wilk test), then the *p*-value in the final column refers to a one-tailed, unequal-variance t-test of the difference in mean vulnerabilities. A one-tailed test was used because the interest is in determining whether the number of errors is greater in the dysarthric group, while there is no interest in distinguishing between the other two possibilities (equal vulnerability and greater vulnerability in the control group). In cases where one or both of sets of vulnerability values were found to be non-normal, the *p*-value refers to the results of a one-tailed Mann-Whitney U test. The alternative hypothesis of this test depends on the distributions of the data. If the two populations can be shown to have distributions of a similar shape, then the test pertains to the difference in the median values. Otherwise, the alternative hypothesis is that there is a 50% probability that the vulnerability rate for a

randomly drawn member of the dysarthric population is greater than the vulnerability rate for a randomly drawn neurotypical speaker. Given the small sample sizes, it was not possible to compare distribution shapes, so the second hypothesis was adopted.

Directional contrast error	*Mean vulnerability rate (%): dysarthria*	*Median vulnerability rate (%): dysarthria*	*Mean vulnerability rate (%): control*	*Median vulnerability rate (%): control*	*p-value*
C1 nasal fronting	0.001	0.00	9.31	0.00	N/A †
/ɛ/ → /ɪ/	12.41	12.41	9.95	6.67	0.299
C2 nasal fronting	10.70	10.51	7.05	5.95	0.154
C2 nasal backing	3.62	2.87	1.81	2.35	0.093
C1 singleton → cluster	2.22	2.23	0.45	0.48	0.001
C1 stop voicing	3.72	1.52	4.97	4.71	0.864
C1 nasal backing	11.78	5.00	4.45	2.63	0.335
/ɪ/ → /i/	11.78	9.03	7.60	5.42	0.252
C2 singleton → cluster	2.96	1.15	0.54	0.48	0.034
C1 cluster → singleton	4.74	1.04	1.62	0.00	0.031
C1 stop devoicing	9.95	6.69	2.59	1.46	0.013
Monophthong → diphthong	1.07	0.71	0.38	0.10	0.011
Diphthong → monophthong	15.03	10.46	4.02	3.92	0.0008

† Significance testing was not possible in this case, since only one error arose in the dysarthric group. However, it is the only error that occurs more often in neurotypical speakers. Therefore, it is shaded in grey (see caption).

Table 5.5. Mean and median vulnerability rates for the most prominent directional contrast errors. The last column shows the result of significance testing, either using either the Student's t-test (above the bold line) or the Mann-Whitney U test (below the bold line). Within each test type, the data are presented in order of decreasing *p*-value. Grey shading means that there is *no* evidence that the contrast error occurs more often in speakers with dysarthria, while unshaded means that there is *strong* evidence that the error is dysarthric ($p < 0.05$).

The vulnerability rates in Table 5.5 should be regarded as approximations, due to the simplifying assumptions made when computing the denominators. Further, the *p*-values should be interpreted with caution, as the calculation of a series of statistical tests increases

the risk of a type I error. The possibility of adjusting the *p*-values was considered. However, standard adjustment methods tend to be overly conservative or require the variables to be uncorrelated, which was not the case in the present analysis. In addition, at the current stage of development of the Dutch dysarthria assessment, type I errors (the identification of a confusion as "dysarthric" when it is actually "normal") would be more acceptable than type II errors. This is because future research (with larger sample sizes) is required to validate the word list and the contrast categories proposed in this thesis, meaning that there will be further opportunities for rejecting categories that are not useful for dysarthria detection. On the other hand, eliminating categories that are in fact dysarthric at this stage of the development of the test would be highly undesirable.

The following contrasts failed to show evidence of being more vulnerable in speakers with dysarthria, implying that, on the basis of this preliminary analysis, these confusions do not arise due to impaired speech production:

- C1: stop voicing, nasal place (both fronting and backing)
- Vowels: /ɛ/ → /ɪ/, /ɪ/ → /i/
- C2: nasal place (both fronting and backing)

5.3.6. Recalculation of thresholds for dysarthria detection

Having established that some of the contrast categories identified in Study 1 are unlikely to be relevant to dysarthria detection, it is worth recalculating word-accuracy scores for both speaker groups, in order to determine the effect of omitting these categories. The following categories were removed: (1) Nasal place errors (either backing or fronting and at both word positions); (2) C1 stop voicing errors (i.e., the perception of either /p/ or /t/ as their respective voiced counterparts); (3) The directional vowel errors /ɛ/ → /ɪ/ and /ɪ/ → /i/. The data of all speakers were inspected to identify transcribed words that *only* contained errors from these categories, including words that contained more than one such error simultaneously (e.g., /zɪŋ/ transcribed as /zin/). A new accuracy metric was then calculated for each speaker in which these words were re-scored as correct. The results are presented in Fig. 5.6a, along with the original data (where no categories were excluded) in Fig. 5.6b. The cutoffs for dysarthria detection, below which a diagnosis of dysarthria would be indicated, are shown for the 95% (blue line) and the 97.5% (orange line) confidence levels (one-sided). These cutoffs were calculated from the normative data, in the same manner as described in Section 5.3.1. For the 97.5% confidence level, the threshold value increases from 83.1% when all errors are included (Fig 5.6b) to 87.5% when the non-

dysarthric errors are excluded (Fig. 5.6a). The corresponding values for the 95% confidence level are 84.4% and 88.5% respectively.

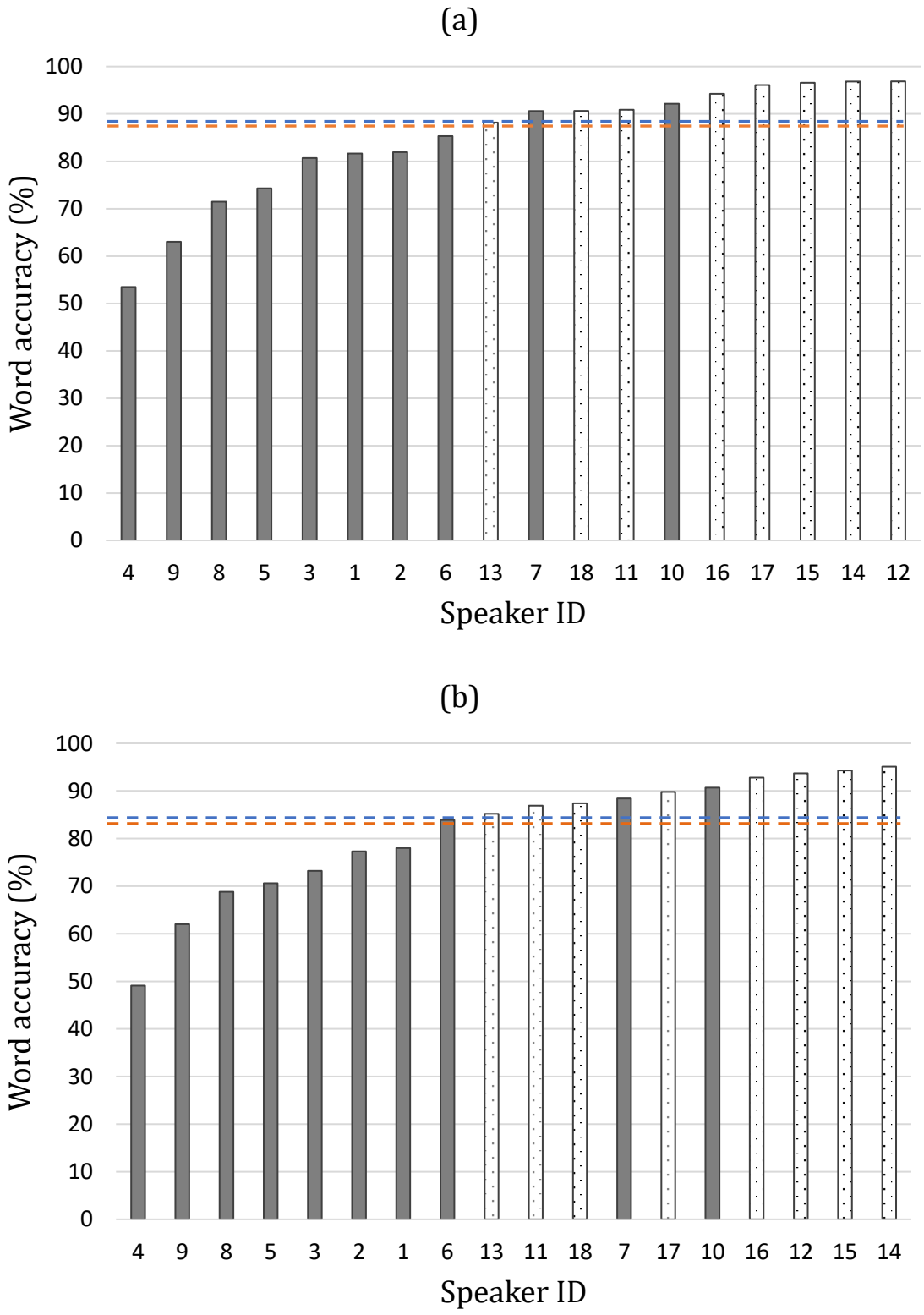


Figure 5.6. Word-accuracy values (a) excluding and (b) including categories that are likely to be non-dysarthric. (Dotted pattern: neurotypical speakers; filled bars: speakers with dysarthria). The blue and orange lines show 95% and 97.5% confidence levels for dysarthria detection, respectively.

The range of word-accuracy scores in the control group after exclusion of the errors classified as "normal" was 88.2% to 96.9%, with a mean (± 1 SD) of 93.8 ± 3.4%. Figure 5.6 reveals that, for a 97.5% confidence level (orange line), one of the speakers who was categorised as *dysarthric* based on his spontaneous-speech intelligibility (S6) has a diagnosis that switches from non-dysarthric to dysarthric when the confusions regarded as "normal" are excluded from the analysis. If, on the other hand, a 95% confidence level is assumed (blue line), then one of the *control* speakers (S13) receives a diagnosis that switches from non-dysarthric – when all errors are counted – to dysarthric when the normal errors are neglected. All other speakers receive the same diagnosis irrespective of both the confidence level for the cutoff and the way in which "normal" errors are treated.

5.4. Discussion

5.4.1. Word-accuracy scores and cutoffs for dysarthria detection

The present findings are consistent with those of previous studies (e.g., Vigouroux & Miller, 2007; Haley et al., 2000) in which 'notable' error rates, i.e., word-accuracy scores lower than about 90%, were observed in some neurotypical speakers. The scores obtained after the removal of errors that were clearly non-dysarthric (word-accuracy range 88.2% to 96.9%, mean value 93.8%) were highly similar to those reported by Haley et al. (2000) for orthographic transcription of the words from Kent et al.'s (1989) assessment (range 88.5% to 97.8%, mean 95.2%). However, it is important to note that the quantitative between-group comparison carried out in the present study (see Section 5.3.5) was rather limited in scope (due to low statistical power), and future research might reveal that additional categories (e.g., /h/ deletion) and/or test stimuli (e.g., /ʃɔu/) should be excluded from the assessment, based on the finding that they yield similar error rates in the two groups. If so, then the word-accuracy scores would increase in both groups, as would the cutoffs for dysarthria detection based on normative data.

In addition to contrast categories that yielded errors in a wide range of neurotypical speakers, there were confusions that were not prominent in the cohort *as a whole*, but were reasonably consistent in particular individuals; examples included /h/ deletion and substitution of word-final /r/ with a fricative. Further research with a larger sample size would be required to produce an accurate estimate of the true incidence of these errors in neurotypical speakers and to try and determine the underlying cause. Contrast confusions in neurotypical speakers are a consequence of one or more of a variety of mechanisms that might include: normal phonological processes (phonemic neutralisations) in a particular accent; confusions between phonemes of high perceptual similarity (especially in the

presence of recording noise); distortions due to assimilatory processes with neighbouring phonemes; unnatural productions that arise in single-word reading (e.g., hyper-articulation); and age-related articulatory imprecision. By separating out these causes, it would become possible to develop an assessment that is targeted to the particular population and has the greatest possible utility in diagnosing both the presence of dysarthria and the vulnerability of specific phonetic contrasts. Ultimately, it may prove necessary to define cutoffs for *individual* contrast categories, such that only error rates above a certain value are considered disordered.

5.4.2. Vulnerability of consonant phonemes

The consonant class most prone to error in neurotypical speakers was nasals, due to the high incidence of 'nasal place' confusions. As mentioned throughout the thesis, nasals are thought to have high perceptual similarity, so this finding is not surprising. The voiceless labiodental fricative also showed high error rates. Due to the small number of tokens tested, this result would need to be corroborated. However, it agrees with the data of Pols (1983), who investigated phonemic confusions in neurotypical speakers under various acoustic-disturbance conditions. In the Pols study, the stimuli consisted of pseudo-words of the form CVCVC embedded in carrier phrases. Speakers were instructed to stress both syllables equally "with perhaps some emphasis on the first syllable". Averaging across all acoustic conditions, /f/ was found to be the second most vulnerable word-initial consonant, mainly due to confusions with /v/. Similarly, Pols found that /v/ was frequently perceived as /f/, as a result of which /v/ was ranked the fourth most vulnerable C1 phoneme. Meanwhile, voice confusions (and confusions in general) were much less common for the alveolar fricatives, making them two of the most robust phonemes. This disparity between alveolar and labiodental fricatives was also observed in the present study.

Other phonemes showed moderate or high error rates for specific individuals only: initial-/h/, initial-/ɣ/ and final-/r/. Future studies should pay particular attention to these phonemes, to determine whether their vulnerabilities are more widespread. For the time being, although a comparison with Pols (1983) is of limited value (due to differences in the phonetic and linguistic context), it is worth briefly mentioning his corresponding findings: word-initial /h/ and word-final /r/ both had an intermediate level of robustness, while word-initial /ɣ/ was highly robust. Further discussion of these phonemes is provided in the following subsection on phonetic-contrast confusions.

Finally, it is worth mentioning the most striking discrepancies between Pols (1983) and the present study. Pols found /w/ to be the most vulnerable of all the word-initial consonants,

while /l/ was the third most vulnerable. In contrast, these two C1 phonemes were perceived with high accuracy in the present study. The discrepancies are likely to be due to differences in study design. The settings used by Pols (especially the noise and the lack of semantic cues) would have had a detrimental effect on the perceptual distinctiveness of phonemes (relative to the present study), and it seems plausible that some phonemes would have been affected to a greater extent than others. The alveolar lateral approximant, for which the most frequent confusions in the Pols study were either /n/ or /r/, is a perceptually weak phoneme that is often difficult to identify on spectrograms. Furthermore, it is prone to confusion with other approximants when judged by speakers of languages that do not contain the phonological contrast in question. Therefore it is unsurprising that in the presence of noise, /l/ would be particularly prone to error. In the case of /w/, the most frequent confusions were with /h/ and /b/. The confusions with /h/ are uninteresting, as listeners use this phoneme as a default response in the absence of sufficient identifying information, as was also noted by Warner et al. (2005). As for the /w/ → /b/ error, it is perhaps relevant that the perceptual distinctiveness of English /w/-/b/ has been observed to increase as the speaking rate decreases (Miller & Baer, 1983). The authors' explanation was that the duration of the initial formant transitions is an important cue to the /w/-/b/ distinction (with syllables beginning with /b/ having shorter transitions), and this difference is exaggerated at slower speaking rates. Assuming these arguments might also apply to Dutch,[5] it is possible that the monosyllabic, single-word reading task used in the present study was conducive to a slower speaking rate than the task used by Pols (1983) and thus the /w/-/b/ distinction remained intact.

5.4.3. Phonetic-contrast confusions

The first point worth noting is that the phonetic-contrast errors perceived in neurotypical speakers were relatively *inconsistent*, especially when compared with the dysarthric group. The average consistency in speakers with dysarthria was 61% (± 9.5%), where a consistent error was defined as a phonetic-contrast confusion that was perceived by at least two listeners (see Chapter 4, Section 4.3.5). In neurotypical speakers, the mean consistency was 43% (± 13%). The difference in mean consistency between the two populations is highly significant (two-tailed $p < 0.004$ in a two-sample t-test). As argued in Chapter 4, a common cause of low consistency is when the speaker's misarticulations mainly consist of distortions. It seems likely that when a neurotypical speaker pronounces a phoneme in a nonideal way, the degree of phonetic distortion is lower than in speakers with dysarthria.

[5] In fact, it seems likely that the durational cue is even more important for Dutch than for English, as /b/ is produced with prevoicing in Dutch, which may increase its perceptual similarity to /w/.

Furthermore, a higher proportion of the confusions observed in neurotypical speakers (relative to speakers with dysarthria) do not arise due to a misarticulation at all, and are purely perceptual in origin. Again, it seems likely that such confusions will be heard with lower consistency than a true articulatory error.

There were two vowel contrasts and two consonant contrasts that were equally common in control speakers as in speakers with dysarthria (see Table 5.5). The vowel errors were both described as unidirectional: /ɛ/ → /ɪ/ and /ɪ/ → /i/. For the /ɛ/ - /ɪ/ pair, this was indeed the case – there was not a single instance of /ɪ/ perceived as /ɛ/ in neurotypical speakers. In the case of /ɪ/ - /i/, however, there were a small number of /i/ → /ɪ/ confusions. As can be seen in Fig. 5.5, the formant values for /ɪ/ and /i/ were found to be very similar in the Antwerp accent in a study from 2002. Verhoeven (2005) has suggested that the /ɪ/ - /i/ distinction might be undergoing a process of neutralisation.[6] The fact that the errors in the present study predominantly occurred in one direction lends support to this theory, as 'random' errors between phonemes of high perceptual similarity ought to be bidirectional (as was the case for nasal place contrasts). The second confusion, /ɛ/ → /ɪ/, involves the same phonetic processes as /ɪ/ → /i/ (raising and fronting) and it occurs in a similar region of the vowel space. Thus the overall picture is reminiscent of other vowel reorganisations, such as the "Great Vowel Shift" of English, which, among other changes, involved the successive raising of three front vowels. The rationale for such patterns (known as chain shifts) is that the movement of one phoneme in acoustic space causes other phonemes to shift in such a manner so as to maintain phonemic differentiation. Although the chain shift argument seems plausible, to the best of the author's knowledge, previous literature has not drawn specific attention to the fact that /ɛ/ is often perceived as /ɪ/ in the Antwerp Dutch accent. Therefore further research is required, involving speakers of different ages, to determine the true extent of this phenomenon. For the time being, it is worth noting that differences in the *distribution* of errors for the two vowel confusions (/ɪ/ - /i/ and /ɛ/ - /ɪ/) are consistent with fact that the /ɪ/ → /i/ confusion is more widely acknowledged in the literature: for /ɪ/ → /i/, confusions were perceived in all eight control speakers, whereas for /ɛ/ → /ɪ/, two speakers did not yield any errors. Furthermore, the distribution of errors was less uniform for the /ɛ/ → /ɪ/ confusion, and in fact, just two speakers (one male and one female) were responsible for 65% of the errors. These discrepancies suggest that, at present, the formant frequencies of /ɛ/ and /ɪ/ are still

[6] Although note that there is also a marked *durational* difference between /ɪ/ and /i/ in the Antwerp accent, or at least there was such a difference at the time of Verhoeven & van Bael's (2002) study.

reasonably well separated for many speakers, while overlaps in frequency-space between /ɪ/ and /i/ are much more common.

The consonant category 'nasal place' was vulnerable at both word positions and there were errors in both directions (i.e., fronting and backing). Place characteristics of nasals are thought to be difficult to perceive (Narayan, 2008; Black, 1969) and identification of nasal place can present a challenge for acoustic classification techniques (Narayan, 2008). Furthermore, in some languages of the world, nasals are essentially placeless and tend to assimilate in place to the adjacent segments. According to Kawahara and Garvey (2014), a possible explanation for the perceptual similarity of nasals is that coarticulatory nasalisation in adjacent vowels may blur the formant transition information required to make place judgements. These authors carried out a series of detailed experiments on word-final nasals and stops produced by English speakers. Their outcome measures included listener judgments of perceptual similarity and phoneme identification using a forced-choice paradigm. They showed that the place contrast is more stable in stops than in nasals and that this relationship holds (a) for a variety of noise conditions and (b) with and without a clear release burst. Regarding the evidence for Dutch, Pols (1983) showed that the place contrast is perceived more accurately in oral stops than in nasal consonants. The instability of nasal place was particularly striking in word-final position, and as a consequence, /n/ and /m/ topped his list of vulnerable word-final phonemes. Note, however, that the Pols confusion matrices were averaged over all noise conditions, and in the text of the article, he states that he did not observe any word-initial confusions between /m/ and /n/ in the zero-noise condition. In fact, in general, the evidence of nasal place vulnerability in the literature seems to be much stronger for word-final than word-initial position. An interesting finding in the present study was that, for word-initial position, speakers with dysarthria yielded just one /n/ → /m/ substitution, but 23 errors in the opposite direction. Neurotypical speakers, on the other hand, yielded an equal number of errors (7) in both directions (note that there were 6 target words beginning with /m/ and 3 with /n/). This between-group difference cannot be obviously explained and it could just be a statistical anomaly. However, it is worth noting that, in common with the dysarthric group in the present study, Pols (1983) found that C1 /m/ → /n/ confusions outnumbered substitutions in the opposite direction, albeit by a much smaller factor (~ 2.5).

The second consonant contrast-error that occurred frequently in neurotypical speakers, which in fact produced higher mean and median vulnerability rates than in speakers with dysarthria (see Table 5.5), was the voicing of phonologically voiceless plosives. Table 5.5 also shows that 'C1 stop voicing' was almost twice as common as 'C1 stop devoicing' in the

control speakers. As discussed below, the *expected* finding in neurotypical speakers, at least from a production perspective, was that devoicing would arise more often than voicing. A two-tailed paired t-test was applied to the vulnerability rates for these two confusions (C1 stop voicing and C1 stop devoicing) in neurotypical speakers, having shown both datasets to be normally distributed, and a *p*-value of 0.083 was obtained. The mean difference in vulnerability (voicing – devoicing) was 2.37% with 95% confidence intervals of [-0.40%, 5.15%]. In other words, it is fairly *un*likely that these data represent a population in which devoicing is more common than voicing. To investigate this unexpected finding, the first step was to check whether the high error rate for C1 stop voicing could have been an artefact of the approximate method used to calculate vulnerability. Specifically, as mentioned in Section 5.3.5, the denominator assumed that *all* words containing the phonemes in question were capable of forming minimal pairs. Inspection of the word list revealed that the proportion of words beginning with /b, d/ that formed minimal pairs with words beginning with /p, t/ was identical to the proportion of /p, t/ words that could be meaningfully perceived with word-initial voicing (both 75%). Therefore, the result does not appear to be artefactual. As mentioned, based on other evidence in the literature, it was expected that 'stop devoicing' would be the more common finding. Firstly, models based on the "difficulty" theory show that the difference between subglottal and supraglottal pressure is unlikely to exceed the assumed threshold for voice initiation prior to the release of an oral stop, and that the most probable scenario is for word-initial voiced stops to be realised as voiceless and unaspirated (Westbury & Keating, 1996). Indeed, Dutch is one of the few Germanic languages that attempts to contrast voiced and voiceless unaspirated plosives (van Alphen & Smits, 2004). English and German, for example, contrast voiceless unaspirated and voiceless aspirated plosives in initial position. Fully voiced initial plosives, meaning that they are produced with a negative voice onset time (VOT) – or, equivalently, with prevoicing – can only arise when certain physiological and aerodynamic conditions are met. Thus, they are considered to be relatively prone to disruption (van Alphen & Smits, 2004). These authors showed, in an experiment with Dutch speakers, that when listeners were asked to choose between the perception of the voiced and devoiced plosive, the error rate on /b/ (i.e., a devoicing error) was almost twice that on /p/ (a voicing error): 11.6% vs. 5.9%.[7] A very similar result was obtained by Pols (1983), who reported 98 instances of /b/ perceived as /p/ but only 46 errors in the opposite direction. In an attempt to understand why these results were not replicated in the present study, a brief comparison

[7] The discrepancy was smaller for the alveolar plosives: 8.1% for /d/ vs. 7.8% for /t/, a finding that was explained in the paper. However, since most of the instances of initial voicing in the current study occurred on /p/, the bilabial plosive is more relevant to the present discussion.

of the acoustic signals of "correct" and "incorrect" productions of /p/ and /t/ was carried out – that is, a comparison of tokens that were perceived correctly, as voiceless, versus those that were transcribed as voiced (by at least one listener). Van Alphen and Smits (2004) showed that prevoicing is the most reliable cue to the voicing distinction in Dutch initial plosives.[8] Therefore, the first step was to determine whether any of the voiceless tokens that were perceived as voiced in the present study showed evidence of prevoicing. This was not the case.[9] Various other cues have been suggested for distinguishing between voiced and voiceless initial stops in Dutch (van Alphen & Smits, 2004), three of which are relatively easy to implement and were therefore investigated here using Praat (Boersma & Weenink, 2018): (1) the duration of the noise burst (shorter for voiced plosives), (2) the value of F_0 immediately after the burst (lower for voiced plosives), and (3) the change in F_0 between the time-point immediately after the burst and the steady-state portion of the subsequent vowel (positive for /b/ - 'rising'; negative for /p/ - 'falling'). For 4 out of the 5 speakers who yielded voicing errors for /p/, these three characteristics were compared between tokens that were heard correctly and tokens that were heard as voiced. No comparison was possible for the fifth speaker, as *all* of his /p/ tokens were heard as voiced by at least one listener. To summarise the findings, and bearing in mind that they are based on a small number of observations, only the temporal change in F_0 showed evidence of being a likely cause: in two of the four speakers, S11 (female) and S16 (male), all of the /p/ tokens that were sometimes heard as /b/ had a rising pitch pattern, while the tokens that were perceived correctly had a falling pattern. For the remaining two speakers, no clear acoustic explanation emerged. However, it is interesting to note that most (83%) of the instances of 'stop voicing' were observed in male subjects. Furthermore, the subject who showed the error most consistently (and who was therefore not eligible for acoustic analysis, as explained above) had the lowest F_0 values of all the male speakers. Therefore an explanation based on pitch seems likely. Clearly, further research would be required to (a) corroborate the finding that the perception of phonologically voiceless stops as 'voiced' is a relatively common occurrence, at least for some (male) speakers of Antwerp Dutch and (b) understand the acoustic correlates of these misperceptions. For the time being, assuming that the phenomenon is real and not artefactual, it is interesting to consider briefly why it might occur. Van Alphen and Smits (2004) noted that despite the importance of prevoicing to the perception of voiced stops in Dutch, 25% of tokens do not show this

[8] As mentioned above, Dutch voiceless plosives are unaspirated. They are also produced with very short voice onset times, compared to English.

[9] Van Alphen and Smits (2004) reported the same negative finding.

feature. As a result, a relatively high proportion of /b/ tokens in their study were misperceived (10% - much higher than the equivalent error rate for the perception of /b/ in English). They propose that this could indicate a phonological change in process: "The potential diminishing of prevoicing may be caused or boosted by the large influence of English on Dutch." Following on from this suggestion, note that the listeners in the current study were mainly young university students with a high level of English and frequent exposure to the language. Therefore, it is possible that when they hear a plosive *without* prevoicing, they are more inclined (compared to previous generations) to consider the possibility that the intended target was voiced. In other words, a reduced tendency to prevoice (a change in *production*) might result in an increase in the proportion of voiceless targets that young people perceive as 'voiced' (a change in *perception*). Further research would be required to determine whether there is evidence of a phonological change of this nature. However, it is an interesting possibility to consider.

In addition to the contrast categories discussed above, which were prominent in a wide range of neurotypical speakers, there were contrast errors that were mainly observed in just one speaker, although often with moderate or high frequency. The first such error is /h/ deletion. De Louw (2016) claims that /h/-dropping is a known feature of some varieties of Belgian Dutch. The fact that it was only consistently observed in one neurotypical speaker has two possible explanations – either the remaining speakers do not exhibit this trait or they suppressed it during the experiment, perhaps due to a tendency to speak more formally and/or hyper-articulate. In any case, a sample size of 8 is not sufficient to yield a reliable estimate of the incidence of /h/-dropping among speakers of Antwerp Dutch. Therefore, future studies with larger sample sizes will be required. Secondly, one of the neurotypical speakers (S13) yielded eleven errors for /ɣ/. It was transcribed as /k/ on 7 occasions; otherwise, it was either replaced by /h/ or deleted. Consider, first of all, the substitution of /ɣ/ with /h/, a confusion that was also observed in dysarthric speakers. In most accents of Belgian Dutch, the velar fricative is articulated as a so-called *zachte* ("soft") *g*, meaning that it is produced further forward in the mouth than in standard Netherlands Dutch (e.g., post-palatal) and with less energy and scrappiness. According to Collins and Mees (2003: p192), in some accents of Belgian Dutch, the velar fricative can sound more like the glottal fricative, perhaps due to this "soft" articulation. Therefore, although this confusion may not be a recognised feature of the Antwerp accent, it may be an increasing trend and/or it may occur in certain individuals. As for the most prominent /ɣ/ confusion seen in S13, /ɣ/ → /k/, which was not observed at all in speakers with dysarthria, further examination (perceptual and acoustic) of these tokens in S13 revealed that they were not strong plosives, but at the same time, there was no clear frication noise and the duration of

the sound would be considered short for a fricative. A larger normative study would be required to determine whether this realisation is a more widespread phenomenon. However, from the present study, it appears that S13 is an outlier and that the velar fricative is highly robust in most individuals. This is consistent with the findings of Pols (1983). As mentioned, the phonetic context investigated by Pols differed from that of the present study. Furthermore, it is likely that his speakers hailed from a part of the Netherlands that uses *harde* ("hard") *g*. Nevertheless, the velar fricative was found to be the second most robust phoneme in word-initial position. When confusions did occur, the most likely substitution was /h/ followed by /k/, thus providing some corroboratory evidence of the confusions observed for S13 in the present study. The final confusion worth discussing is word-final /r/ → fricative, which was observed in Speaker 18. In fact, this individual yielded 22 out of 70 errors on word-final /r/, but the majority (15) consisted of confusion of /r/ with a fricative, either /x/ or /s/. Although /r/ is generally realised as an alveolar trill in the Antwerp accent, it has numerous allophones across Belgium and the Netherlands, including uvular pronunciations that can sound much like /ɣ/ or /x/. Listeners are familiar with these allophones, so it is unsurprising that /r/ may sometimes be perceived as /x/. The confusion with /s/ is more surprising and, given that the speaker also yielded other contrast errors for word-final /r/, the overall picture suggests that this speaker either had a highly unusual production of this phoneme or perhaps even some degree of speech impairment.

5.5. Summary

The purpose of this study was to acquire normative data to provide a context for interpreting the results in Study 1. Firstly, it was shown that a number of phonetic contrasts do not show evidence of being more vulnerable in speakers with dysarthria than in neurotypical speakers: the voicing of word-initial stops, nasal place confusions (at both word positions), and the vowel substitutions /ɛ/ → /ɪ/ and /ɪ/ → /i/. Having identified confusions that are unlikely to be indicative of dysarthria, the errors observed for these categories were removed from the analysis, and the remaining normative data were used to calculate threshold word-accuracy values below which an individual would receive a diagnosis of dysarthria. The cutoffs for the 95% and 97.5% confidence levels were 88.5% and 87.5%, respectively.

The subsequent study (Chapter 6) was designed to provide further context for interpreting the findings of Study 1. It compares the phonemic and phonetic-contrast errors perceived using orthographic transcription with those obtained in a closed (four-alternative forced

choice) response mode. The advantage of the closed response mode is that the confounding effect of functional load is, in principle, eliminated. Therefore, it becomes possible to differentiate between errors that were prominent in Study 1 due to high functional load and errors that reflect an important production impairment. Furthermore, a comparison of the free- and forced-response modes will yield valuable information about whether these two techniques provide qualitatively different information (over and above the aforementioned differential effect of functional load). Such information would yield both methodological and theoretical insights concerning the interaction between production and perception.

6. Study 3: Multiple-choice identification of phonetic-contrast errors in speakers with dysarthria

6.1. Research questions and hypotheses

The view taken in this thesis is that phonetic-contrast analysis should ideally be implemented using a forced-response format. However, the latter response mode is less representative of how speech is encoded in real-world communication, and as discussed in Chapter 2, a number of mechanisms can be imagined by which the method might introduce bias into the listener responses. While a few studies have calculated word-accuracy scores for open and closed response modes, there is very little prior research that compares the *nature* of the information yielded by each method, i.e., whether they produce similar profiles of articulatory errors.

Following the literature review, four objectives were identified in relation to the present study (see Chapter 2, Section 2.5.3). The first was to test the following hypothesis:

> *Intelligibility metrics derived from single-word reading are higher for the forced-response mode than the free-response mode.*

The second objective could not be investigated using a hypothesis or a testable question, and was therefore expressed in broad terms:

> *What is the consistency of the relationship between word-accuracy scores for the free- and forced-response modes?*

The third objective related to the level of inter-rater agreement for phonetic-contrast errors identified using the closed method. As stated in Section 2.5.3, since it was not possible to calculate a metric of inter-rater reliability that is directly relevant to the outcome measure, this objective was expressed as follows:

> *To obtain preliminary data regarding inter-rater reliability in a forced-choice single-word intelligibility test of Belgian Dutch speakers with dysarthria.*

The fourth objective was to test the hypothesis that there is a positive association between the profile of phonetic-contrast errors yielded by the two techniques (open and closed):

> *The degree of correlation between the ranked errors yielded by the two response modes exceeds zero, both in the case of individual speakers, and when error ranks are summed over the cohort.*

In addition to the four objectives identified in the literature review, a set of hypotheses emerged from Chapter 4 (see Table 4.9) regarding the predominant error directions that

would be observed in the multiple-choice mode, at least for some of the consonant contrast categories. These predictions are reproduced in Table 6.1. The table also serves as a reminder of the final set of consonant contrast categories chosen for testing in the multiple-choice study (the reasoning for these choices was provided in Table 4.9). Recall that the vowel confusions could not be reduced to phonetic-contrast categories, so vowel distractors were chosen to reflect the vowel substitutions most commonly observed.

Consonant contrast category	*Predicted error direction*
Word-initial voice (fricative and stop)	Devoicing
Fricative place	Backing
Stop place	None
Nasal place	None
Stop vs. fricative	None
Stop vs. nasal	Denasalisation
Initial /h/ vs. null	/h/-deletion
Initial consonant vs. null	Null → consonant
Final consonant vs. null	Null → consonant
Initial cluster vs. singleton	None
Final cluster vs. singleton	None
/r/ vs. /l/	None
/r/ vs. fricative	None

Table 6.1. List of consonant contrast categories tested in the multiple-choice study along with predictions for the predominant error direction that would be observed. "None" implies that there was no clear evidence for making a prediction about directionality.

6.2. Method

Multiple-choice listening sessions were carried out using the single-word utterances of speakers with dysarthria only. This is because a preliminary analysis of the data showed that the forced-response format substantially reduced the number of reported errors compared to orthographic transcription. Given that the control subjects already yielded few errors in the free-response mode, further assessment using the forced-response paradigm would not have been a worthwhile use of resources. In fact, the number of listeners was insufficient to analyse the data of all speakers with dysarthria, given that three listeners was considered the minimum number that should judge each word uttered by each speaker. Therefore, the sample size had to be reduced to eight speakers in the present study. It was decided that the two speakers with ALS (one of whom had an unconfirmed diagnosis) would be omitted from the analysis. These speakers can be

considered as outliers in the sense that the onset of their dysarthria had been very recent. At the point in time when the listening data were collected, they had not received any formal diagnosis, nor undergone any neuroimaging.

Three distractors were chosen for each of the 116 words assessed in the multiple-choice study. Thus the listeners had the choice of four words in total, as was the case in Kent et al. (1989). The distractors involved an error corresponding either to one of the consonant contrast-categories defined in Table 6.1 or to a vowel confusion that was common for the target in question. For some of the target words, additional, uncommon contrast categories were included among the distractors, despite the fact that these confusions were not explicitly tested in the multiple-choice study (meaning that they were not tested on a sufficient number of occasions to measure the error rate with reasonable reliability). Examples included /v/ → /w/, /j/ → fricative and /ʏ/ → /ø:/. This strategy was adopted when (a) the contrast in question was observed frequently for the given target word in the orthographic-transcription study and (b) it was difficult, when confined to just the common contrast categories, to devise three distractors that were all considered to be "worthwhile" (meaning that evidence from the orthographic-transcription study suggested that they might be selected). For further details about how the distractors were chosen, the reader is referred to Chapter 3 (Section 3.4.3). The list of distractors is provided in Appendix 3.

The remainder of this section describes the data analysis methods used to address each of the objectives listed in Section 6.1. The first objective was to test the hypothesis that intelligibility metrics derived from single-word reading are higher for the forced-response mode than the free-response mode. This was assessed by applying a two-tailed paired t-test to the word-accuracy scores (having first confirmed that the data were normally distributed using the Shapiro-Wilk test). A two-tailed test was appropriate, even though a particular directionality was expected, because (a) there would have been equal interest in detecting a difference in the opposite direction (i.e., a higher accuracy for the open response mode), and (b) if the data had shown a difference in the opposite direction, it would not have been justified to attribute that difference to random sampling, i.e., there was some prior evidence in the literature (see Chapter 2, Section 2.1.6) that a higher accuracy might be observed for the open mode.

The second objective was to obtain information about the consistency of the relationship between the accuracy scores for the two response modes. As explained in Chapter 2 (Section 2.5.3), there were no specific hypotheses associated with this objective, as there are no clear guidelines for defining a level of consistency that would be considered "acceptable" in the sense that the two response modes can be considered to provide the

same qualitative outcome. Consistency was analysed in three ways. Firstly, the difference in accuracy between the two response modes was calculated for each individual speaker and for the cohort as a whole (i.e., mean difference ± 1 SD). Secondly, the two sets of accuracy scores were plotted on a single bar chart, allowing for visual appreciation of how the rankings of the participants differ in the two response modes. Thirdly, Pearson's *r* (one-tailed) was used to quantify the degree of correlation between the two sets of scores, having first ensured that the scores pass the Shapiro-Wilk test for normality. A one-tailed test was justified because it was considered highly implausible that there would be a significant *negative* correlation between word-accuracy scores in the two response modes (i.e., one that was not a consequence of random sampling).

The third objective was to obtain a preliminary impression of the level of inter-rater agreement for the forced-choice mode. This was achieved by calculating the kappa statistic for each speaker, based on the level of agreement between the listeners for each test item. This yielded an index between 0 and 1 representing the overall level of agreement for the speaker, relative to that which would be expected by chance.

The final set of analyses examined the level of agreement between the error profiles for the two response modes. Separate analyses were carried out for consonant and vowel error profiles, owing to the fact that consonant errors were reduced to phonetic-contrast categories, while vowel errors were reported as specific phonemic substitutions. Therefore, it seemed likely that the level of correlation between the two response modes would be higher for consonants, where each contrast error can correspond to a larger set of phonemic substitutions. To examine the extent to which the two modes gave converging assessments of the types of error made, the error index for each of the response modes was ranked from low to high across the different contrast categories for a given speaker. Rank ordering was employed because the error scores for free and forced-choice recognition cannot be directly compared in any meaningful way, as discussed below (see Section 6.3.3). Pearson's *r* was used to examine the degree of correlation between the two sets of rankings (free and forced) for each speaker. In addition, the ranks for each contrast category in each response mode were summed across all eight subjects. This enabled calculation of two overall measures of correlation between the response modes, one for vowel errors and one for consonant errors. One-tailed significance levels were reported for the same reason as that given above with respect to accuracy scores: it was considered highly implausible that there would be a significant negative correlation between error rankings in the two response modes. Therefore, the null hypothesis was that the two sets of rankings were independent, and the alternative hypothesis was a correlation coefficient in excess of zero.

6.3. Results

6.3.1. Word-accuracy scores and intelligibility rankings

Table 6.2 presents the word-accuracy score for each speaker and compares with the corresponding values from the orthographic-transcription study. As in previous chapters, owing to the fact that the number of *independent* listeners who judged each speaker was not constant, word accuracy was calculated as the number of words transcribed correctly as a percentage of the total number of observations. It can be seen that accuracy scores are higher in the multiple-choice (MC) mode for all speakers. The absolute difference ranges from 4.3% to 24.6%, with a mean (± 1 SD) of 13.1% ± 6.9%. The two sets of scores are significantly different: $t(7) = 5.38$, $p = 0.001$ in a two-tailed paired sample t-test. The 95% confidence interval for the difference is [7.31%, 18.79%], lending strong support to the hypothesis that higher scores are obtained in the forced-response mode. Comparison of word accuracy in the free-response mode (Column 4) with increase in accuracy between the two response modes (Column 6) shows that, in general, the increase in accuracy is greater for speakers of lower intelligibility. Indeed, Pearson's r between Column 4 and Column 6 is -0.80 ($p = 0.02$, two-tailed). This is unsurprising, as there is a ceiling effect for speakers of higher intelligibility. Figure 6.1 shows the data of Table 6.2 presented as a histogram. This makes it easier to appreciate how the intelligibility rankings of the participants differ in the two response modes. It is immediately apparent that although the speakers are not ranked in precisely the same order, the two sets of scores follow a very similar trend; indeed, Pearson's r reveals a strong positive correlation of 0.86, $p = 0.003$ (one-tailed).

In addition to examining the effect of the MC mode on overall word accuracy, it could be of interest to determine whether the relative accuracies of the three segments (C1, V and C2) have changed. Recall that in the orthographic-transcription study, the highest accuracy was for final consonants, while the lowest accuracy occurred for vowels (referred to herein as an *accuracy pattern* of C2 > C1 > V). A repeated-measures ANOVA followed by post-hoc pairwise comparison tests revealed that the only significant difference was between V and C2.[1] Segmental accuracies were also calculated for the eight speakers in the multiple-choice study by computing the number of correct realisations of each segment as a percentage of

[1] This calculation was repeated for the reduced sample in the present study (n = 8). The main effect of word segment remained significant, $F(2,14) = 4.41$, $p = 0.033$, while the post-hoc comparison between V and C2 was weakly significant, $p = 0.068$ (Bonferroni corrected).

the number of occasions on which the segment appeared in the word list.[2] The mean accuracy values (± 1 SD) were as follows: C2 = 95.7 ± 4.6%, C1 = 94.8 ± 3.1%, and V = 93.8 ± 3.8%. Thus the average accuracy pattern remained the same as for the free-response study: C2 > C1 > V. On this occasion, however, a repeated-measures ANOVA showed no significant effect of word segment: $F(2,14) = 0.72$, $p = 0.50$. This is unsurprising, as the segmental accuracy scores for the MC study are approaching the ceiling of 100% (in fact, one speaker achieved 100% accuracy at C2 position). Therefore, it is less likely that significant differences between segments will be observed. There were few differences between the two studies at the level of the individual speaker; i.e., most speakers showed the same (or a very similar) accuracy pattern in the two response modes.

ID (M/F)	*Diagnosis*	*# words*	*Word accuracy: free (%)*	*Word accuracy: forced (%)*	*Diff. (%)*
2 (F)	CVA (suspected to be in brainstem)	106	77.3	85.5	8.2
3 (M)	Medulloblastoma / surgical damage (left cerebellum)	106	73.2	83.0	9.8
4 (F)	Hemangioblastoma / surgical damage (fourth ventricle)	114	49.1	73.7	24.6
5 (M)	Progressive cerebellar atrophy	87	70.6	90.8	20.2
6 (M)	CVA (right cerebellum)	116	83.9	93.4	9.5
7 (F)	CVA (pons / cerebral peduncle – left side)	114	88.4	92.7	4.3
8 (M)	Cortical watershed CVA (PCA / MCA)	116	68.8	85.9	17.1
9 (M)	CVA (left cerebellum)	94	62.0	72.7	10.7
Mean across cohort (± 1 SD)			**71.7 ± 12.4**	**84.7 ± 8.0**	**13.1 ± 6.9**

Table 6.2. Comparison of word accuracies in the free- and forced-response modes. The final column ("Diff") shows the absolute increase in accuracy between the two response modes.

[2] Unlike the segmental accuracies calculated for the free-response study, this is *not* a precise measure of the vulnerability rate. To calculate the latter, one would need to consider the number of occasions on which each segment was actually tested by the MC foils. Since some participants had missing data, this calculation would have been laborious and was considered not to be worthwhile given that the segmental accuracies were close to ceiling in the MC mode.

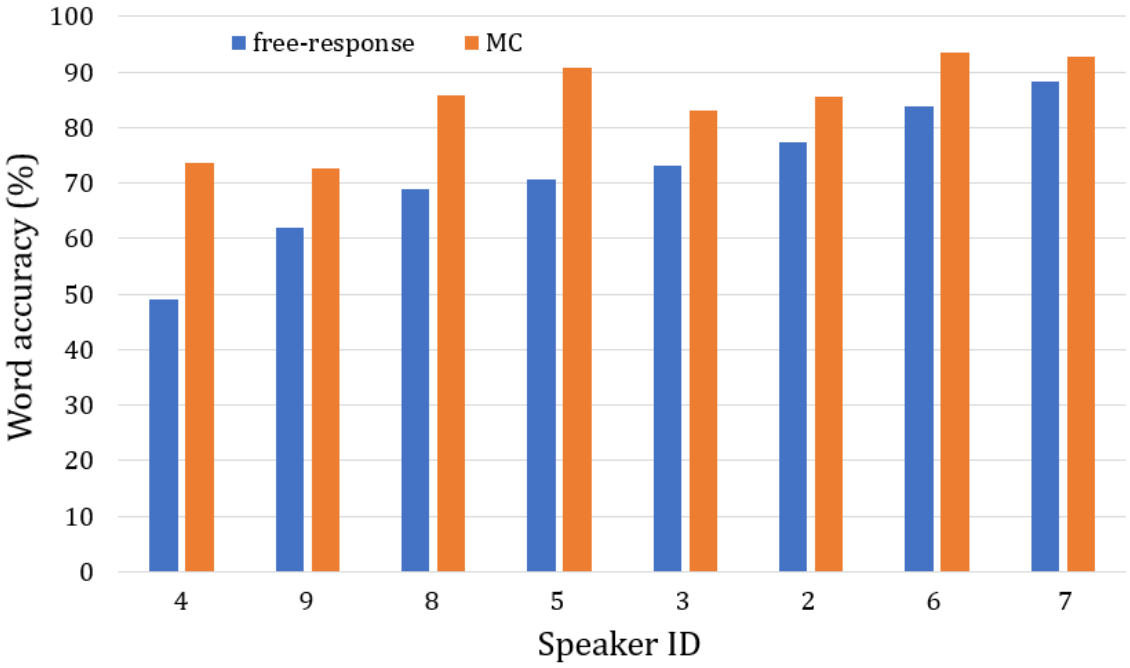


Figure 6.1. Comparison of word accuracies in the free-response and multiple-choice (MC) studies. The data are presented in order of increasing word accuracy for the free-response mode.

6.3.2. Inter-rater agreement

Before examining the qualitative data, an analysis of inter-rater agreement for these data is presented. As explained in Chapter 2 (Section 2.5.3), this metric should be considered preliminary in nature (with regard to assessing the likely clinical value of the technique) due to (a) the suboptimal listener characteristics, (b) the early stage in the development of the proposed dysarthria assessment, and (c) the fact that the reliability metric had to be calculated on a single-word basis rather than for the final outcome measure (the phonetic-contrast error profile). This third issue also means that the reliability data have limited relevance to the present study, as most of the findings discussed in this chapter relate to phonetic-contrast error profiles. Inter-rater reliability would be higher for error profiles than for individual test items, due to the fact that different listeners may yield similar error rates for a given contrast category, but distributed differently over the target words.

In contrast to the free-response mode, where a novel metric of inter-rater agreement ("consistency") had to be devised, it was possible to invoke standard methods of measuring inter-rater agreement in the MC study due to the fact that the listeners' responses were constrained. The appropriate metric is Fleiss' kappa, as (a) it can be used with categorical data (in this case, the four possible responses), (b) it allows for more than two raters, and (c) different items may be rated by different groups of individuals, provided the number of raters per item remains the same. The kappa statistic was calculated for each speaker using

the equation given in McHugh (2012). As shown in Table 6.3, for six out of eight speakers, kappa was determined to be between 0.41 and 0.60, which is generally interpreted as moderate agreement. For Speaker 7, the kappa value of 0.35 corresponds to fair agreement. The lowest level of inter-rater agreement was obtained for Speaker 6 (a kappa of 0.17, generally interpreted as slight agreement). Table 6.3 also shows the consistency metric used to measure inter-rater agreement in the free-response study. It represents the proportion of phonetic-contrast errors that were heard by at least two listeners. Only the *errors* were included in this metric, so the fact that raters effectively agreed on phonemes that they all heard correctly was not taken into account. Since the two metrics are quite different in meaning, and since they apply to different data (free- vs. forced-response), one would not necessarily expect perfect correlation between them. However, Pearson's correlation coefficient was in fact relatively high: $r = 0.76$, $p = 0.03$ (two-tailed). Table 6.3 further indicates that Fleiss' kappa shows a moderate negative correlation with word accuracy (Pearson's $r = -0.63$), although this result just failed to meet statistical significance ($p = 0.10$, two-tailed). As discussed in Chapters 4 and 5, there is at least one mechanism that would result in a negative correlation between intelligibility and inter-rater agreement, namely that speakers who are more severe are likely to yield a greater proportion of substitution (as opposed to distortion) errors, and these tend to be heard more consistently by different listeners. However, since Fleiss' kappa also rewards agreement on *correct* items (unlike the consistency measure in Chapter 4), there is at least one mechanism acting in the opposite direction, namely that there will be higher reliability for speakers who are more intelligible, as they yield a greater proportion of correctly perceived targets.

Speaker ID	*Fleiss' kappa in MC study*	*Consistency in free-response study (%)*	*Word accuracy in MC study (%)*
2	0.49	66.7	85.5
3	0.51	69.8	83.0
4	0.41	68.9	73.7
5	0.45	63.6	90.8
6	0.17	55.2	93.4
7	0.35	65.6	92.7
8	0.42	57.8	85.9
9	0.56	69.4	72.7
Pearson's r (p-value)		0.76 ($p = 0.03$)	-0.63 ($p = 0.10$)

Table 6.3. Fleiss' kappa for the multiple-choice study. The consistency metric used to measure inter-rater agreement in the free-response mode is shown for comparison. The final row shows the correlation between consistency and Fleiss' kappa and between MC word accuracy and Fleiss' kappa.

6.3.3. Consonant contrast errors

The vulnerability rates for the consonant contrast categories, averaged over the cohort, are shown in Fig. 6.2. This graph was produced by dividing the number of observed errors by the number of occasions on which the contrast was tested, and then averaging over the cohort. Similar histograms have been produced by previous authors (e.g., Kent et al., 1990; Blaney & Hewlett, 2007; Whitehill & Ciocca, 2000b), although these studies calculated vulnerability rates for the category as a whole (i.e., irrespective of the error direction), whereas Fig. 6.2 displays separate vulnerability rates for each direction by means of stacked columns. Two of the consonant categories that were tested in the MC study (see Table 6.1) are not shown in Fig. 6.2 – initial consonant vs. null and final consonant vs. null – as these categories did not produce *any* errors across the entire set of speaker-listener observations. In the case of 'initial consonant vs. null', this is broadly consistent with the finding for orthographic transcription, as the category did not produce appreciable errors. However, the category 'final consonant vs. null' was the third most prominent C2 error in the free-response mode.

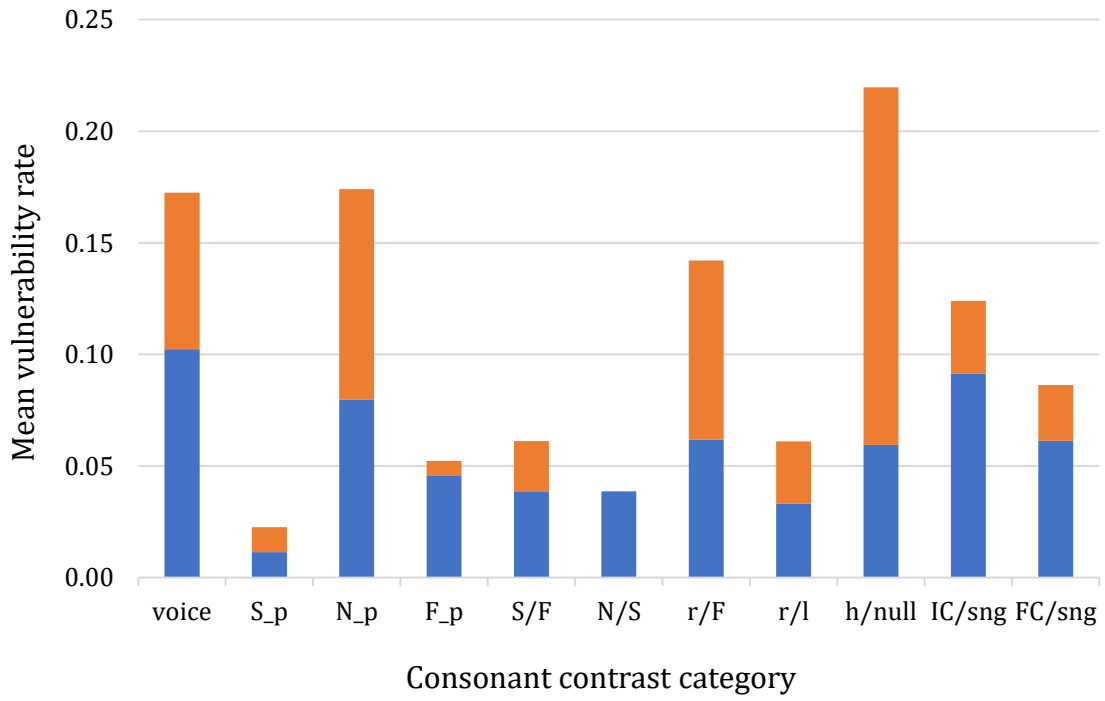


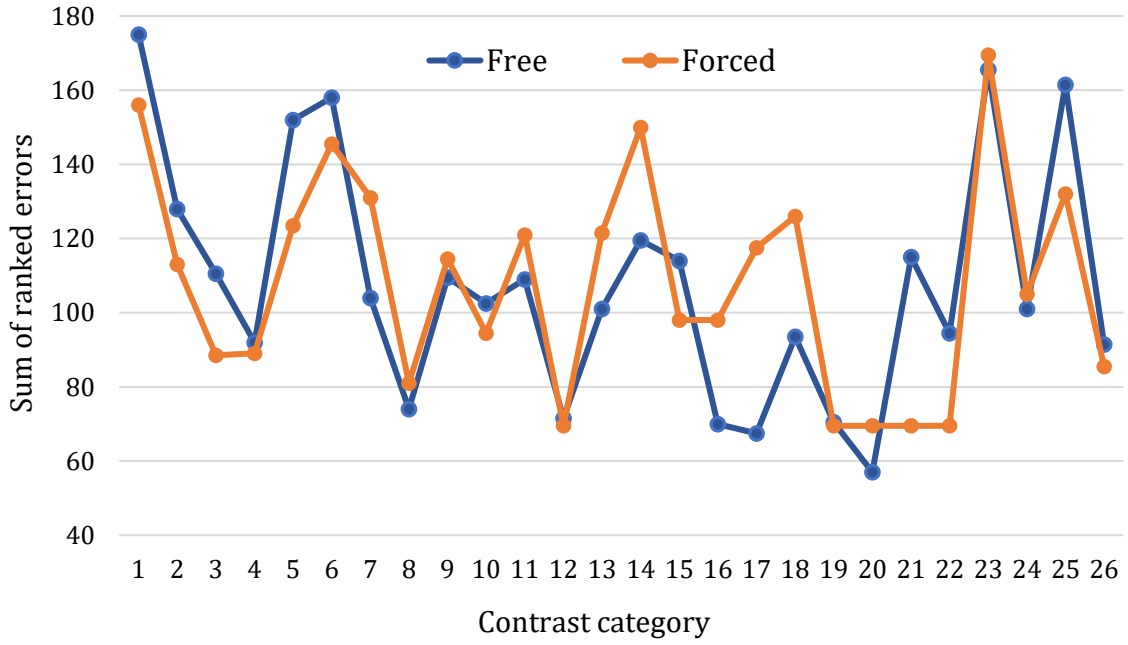
Figure 6.2. Mean vulnerability rates for the consonant contrast categories in dysarthric speakers. Blue (orange) refers to the error direction: devoicing (voicing) for stop and fricative voicing errors (voice); backing (fronting) for stop, nasal and fricative place errors (S_p, N_p and F_p); stop → fricative (fricative → stop) for the category S/F; nasal → stop (N/S); /r/ → fricative (fricative → /r/); /r/ → /l/ (/l/ → /r/); and addition (deletion) for /h/ vs. null (h/null), initial cluster vs. singleton (IC/sng), and final cluster vs. singleton (FC/Sng).

The first observation to be made from Fig. 6.2 is that the predictions regarding the predominant error *directions* (see Table 6.1), for the four contrast categories that yielded errors[3] (voice, F_p, N/S and h/null), were accurate. This demonstrates that, at least for some categories, it is possible to predict the error direction in a forced-response mode based on a combination of the error rates in a free-response mode and approximate knowledge about the functional loads of the two directional confusions. The next question is whether the two response modes yield a similar *error profile*, meaning the relative importance of the different error categories, both for individual speakers and for the cohort as a whole. The remainder of this subsection addresses this question.

The multiple-choice study explicitly assessed 26 directional consonant contrast categories. To examine the similarity between the error profiles in the two response modes, each of the raw error indices for these categories was ranked from low (1) to high (26) on a within-subject basis. Rank ordering was employed for several reasons. Firstly, the raw scores for free and forced-choice recognition are conceptually different. The score for the open mode is effectively a *count* variable (the number of errors observed for a given category), while the score for the forced-choice mode is a *proportion* (the ratio of the number of observed errors to the number of potential errors). Secondly, the use of ranking ensures that the variances of each person's scores are almost equal (because everyone uses the range 1 to 26, except for minor variations due to ties). Comparing populations in which subjects vary substantially in terms of the variance in their scores across the categories (which would have been the case here when using raw scores) is problematic for parametric tests. Thirdly, the scores for most speakers were not normally distributed across the categories, especially in the free-response mode where a large proportion of categories yielded a score of zero. For all these reasons, ranking the error metrics within each mode was the only legitimate way of assessing whether the categories were similarly ordered with respect to their likelihood of generating errors across the two response modes. Each category was represented by the sum of the ranks across all subjects (Fig. 6.3). Note that the minimum *theoretical* value for the sum of the ranks (i.e., when there are no errors for the category in question – which only occurred for the forced-choice mode, e.g., for categories 19-22) is higher for the forced mode than the free mode. This is because the former yields a larger number of zero-error categories per speaker, such that the mean rank of these categories is higher. Therefore, it would be imprudent to compare the *values* of the rank totals in the two modes for any given category too closely; rather, the goal is to assess the extent to

[3] The remaining two categories for which an error direction was predicted – initial consonant vs. null and final consonant vs. null – did not yield any errors.

which the two modes agree in terms of their *ordering* of the categories. Using the rank totals, a strong correlation was found between the two response modes: $r = 0.735$, $p < 0.001$ (one-tailed). Figure 6.4 displays the scatter plot for the rank totals yielded by the two modes, along with the best-fit linear trend. There is a hint in these data of better agreement between the two methods for contrast categories that generate higher error rates, although the data do not have sufficient power to test this apparent trend statistically. Note also that the trend may arise for artefactual reasons, such as the fact that, as mentioned, zero-error categories receive a higher ranking in the forced mode than in the free mode.



1	devoicing	2	voicing
3	stop backing	4	stop fronting
5	nasal backing	6	nasal fronting
7	fricative backing	8	fricative fronting
9	stop to fricative	10	fricative to stop
11	nasal to stop	12	stop to nasal
13	/r/ to fricative	14	fricative to /r/
15	/r/ to /l/	16	/l/ to /r/
17	/h/ addition	18	/h/ deletion
19	initial consonant addition	20	initial consonant deletion
21	final consonant addition	22	final consonant deletion
23	initial singleton to cluster	24	initial cluster to singleton
25	final singleton to cluster	26	final cluster to singleton

Figure 6.3. Sum of ranked errors for the consonant contrast categories as assessed via the forced- and free-response modes. The category codes are shown in the table beneath the figure.

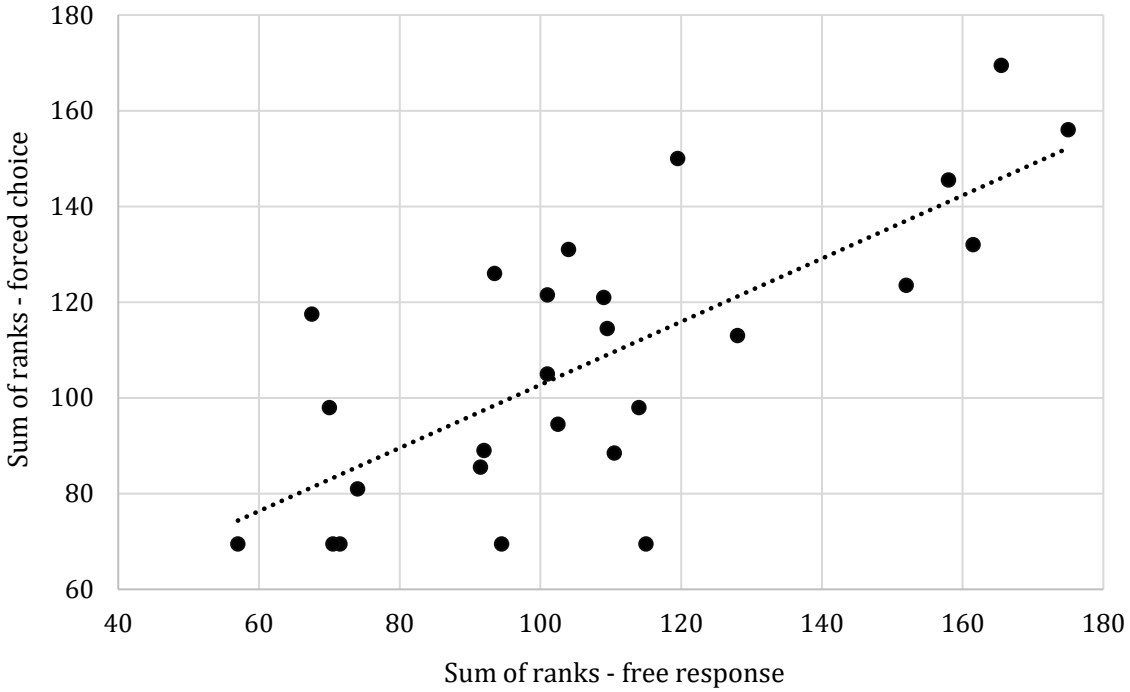


Figure 6.4. Relationship between the total sum of the ranks for the 26 consonant categories.

The top six errors in the free-response mode (in order of decreasing vulnerability) are as follows: (1) devoicing, (2) initial singleton to cluster, (3) final singleton to cluster, (4) nasal fronting, (5) nasal backing and (6) voicing. The equivalent rankings in the forced-choice mode are somewhat different: (1) initial singleton to cluster, (2) devoicing, (3) fricative to /r/, (4) nasal fronting, (5) final singleton to cluster and (6) fricative backing. Some of these differences may arise due to functional load considerations. For example, the categories pertaining to specific consonant classes provide fewer opportunities for errors in the free-response mode than some of the more generic categories (e.g., initial singleton to cluster or voicing / devoicing). Therefore, it is unsurprising that there are two fricative categories (fricative to /r/ and fricative backing) that appear among the top six consonant errors in the forced-choice mode (where the role of functional load is reduced if not eliminated), but not in the free-response mode. However, there are a number of differences between the two response modes that cannot easily be explained by the differential effect of functional load. Possible explanations for such differences are provided in the Discussion (see Section 6.4.2).

Thus far, the findings are reasonably encouraging; in general, the two response modes order the consonant contrast categories in a similar way using aggregate error metrics across subjects. As stated above, when using the rank totals, a strong correlation is found between the two response modes: Pearson's $r = 0.735$, $p < 0.001$ (one-tailed). When discrepancies do arise, they can often be explained by considering the differential effect of

functional load in the two response modes. As discussed below (see Section 6.4.2), it is unlikely that differences arising due to functional load will present a problem for data interpretation in the long term. However, an important clinical question is whether strong correlation between the response modes is also observed at the *individual* level. Unsurprisingly, the correlation between the rank ordering is weaker when it is based on data from individual subjects (see Table 6.4). However it remains statistically significant in every case, except for S7 where *r* falls to 0.341 (p = 0.044, one-sided).

Speaker ID	*Pearson's r*	*p (one-tailed)*
S7	0.341	0.044
S6	0.623	0.000*
S2	0.569	0.001*
S3	0.544	0.002*
S5	0.704	0.000*
S8	0.572	0.001*
S9	0.715	0.000*
S4	0.582	0.001*

Table 6.4. Pearson's *r* for consonant error rankings in the two response modes for each speaker. The speakers are displayed in order of decreasing intelligibility (based on the free-response study). The *p*-values marked with an asterisk are significant assuming a Bonferroni-corrected alpha level of (0.05 / 8) = 0.0063.

To gain insight into the variability in error profiles among speakers, Fig. 6.5 shows the distribution of vulnerability rates (i.e., the number of observed errors as a proportion of the number of potential errors) across the cohort of speakers for eight of the consonant contrast categories. Distributions for the remaining three categories that yielded errors were considered less informative and are not shown: (1) 'stop place' did not yield appreciable errors, as can be seen in Fig. 6.2; (2) 'nasal place' confusions are equally common in neurotypical speakers (see Chapter 5); and (3) '/h/ vs. null' errors are problematic for a number of reasons. In the case of /h/-deletion, while this confusion had a moderate level of vulnerability (see Fig. 6.2), the errors were predominantly due to one speaker, who yielded a vulnerability rate of 0.95. Furthermore, /h/-dropping was a relatively consistent process in one *neurotypical* speaker and could prove to be a feature of the Antwerp accent (this would need to be confirmed in a study with a larger control

sample). Regarding /h/-addition, Fig. 6.2 reveals that this error was perceived on a greater proportion of occasions than /h/-deletion. However, /h/-addition was only tested on four occasions in the forced-choice study; thus, the finding is not considered to have high reliability. In fact, it arises from just six errors across all speaker-listener observations. Therefore further research would be required to determine the vulnerability of /h/-addition in Belgian Dutch speakers with dysarthria.

Due to the low sample size, as well as the limited level of inter-rater agreement, any findings that emerge from Fig. 6.5 regarding how the error rates vary across the cohort should be interpreted with caution. Nevertheless, it is worth summarising the most striking observations, which could be used as a springboard for future research. Firstly, it can be seen that three of the error categories arise in all or most speakers: voice, initial cluster vs. singleton and /r/ vs. fricative. Of these error categories, the first two were also relatively prominent in neurotypical speakers (albeit in a free-response paradigm); thus, it is unsurprising that these categories are found to be vulnerable across all speakers with dysarthria. Furthermore, in the case of 'voice', the vulnerability rate does not appear to increase with speaker severity (from left to right in Fig. 6.5), which would be consistent with the suggestion that the perceived error is not purely "dysarthric". A lack of apparent correlation with overall intelligibility for 'voice' can also be regarded as consistent with van Nuffelen et al.'s (2009b) predictive model of phoneme intelligibility (assessed using the NSVO), where neither voicing nor the lack of voicing emerged as an important phonological feature. With the exception of voice, most of the categories in Fig. 6.5 show a vulnerability rate that increases with speaker severity.[4] In general, such behaviour is to be expected in a study where overall intelligibility is calculated from the same data as those used to determine error rates for specific categories. Nevertheless, based on previous research using phonetic-contrast analysis (see Chapter 2, Section 2.3), including in languages other than English, contrast categories tend to differ in terms of the extent to which such correlation is observed. It could be informative to examine such trends for Belgian Dutch; however, this would require data from a larger sample size.

[4] This statement is based on visual inspection. Due to the small sample size, and the low-moderate inter-rater agreement, it would not be worthwhile examining these trends statistically.

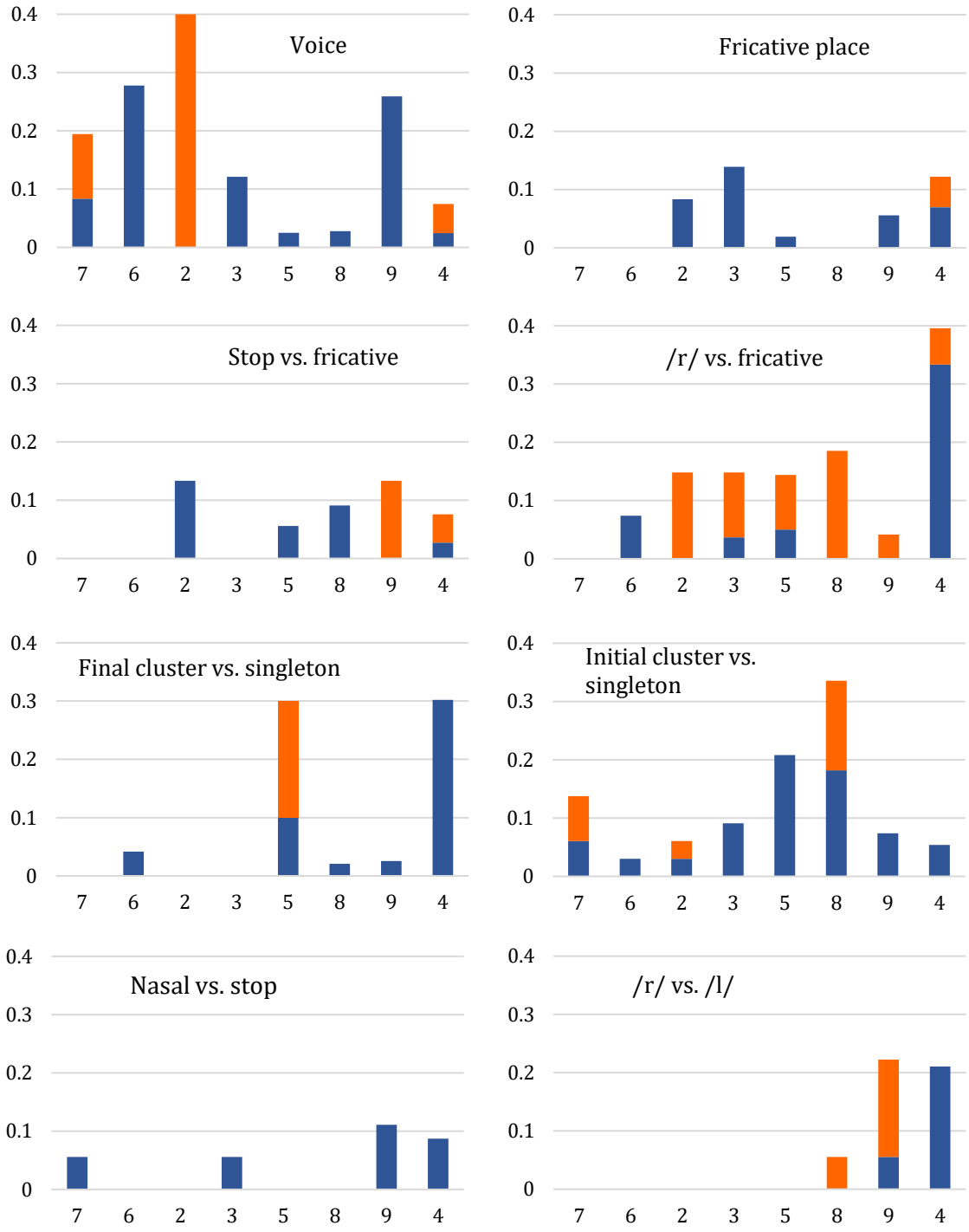


Figure 6.5. Vulnerability rates (y-axis) for individual speakers (x-axis) for eight of the consonant contrast categories. The speakers are presented in order of increasing severity from left to right (i.e., in order of decreasing word accuracy in the free-response mode). Blue (orange) shading refers to the error direction: devoicing (voicing) for the category 'voice'; backing (fronting) for fricative place; stop → fricative (fricative → stop); /r/ → fricative (fricative → /r/); addition (deletion) for final cluster vs. singleton and initial cluster vs. singleton; nasal → stop; and /r/ → /l/ (/l/ → /r/).

6.3.4. Vowel confusions

Given that vowel confusions could not be consolidated into phonetic-contrast categories (with the exception of monophthong vs. diphthong), the vulnerability rates for vowel contrasts were of limited reliability. For example, the category /ɔu/ - /œy/ was capable of generating a maximum of nine errors (three per listener: two in the direction ɔu → œy and one in the opposite direction). This number was even lower when one of the relevant target words was missing for a given speaker. Thus, it was decided that a profile equivalent to Fig. 6.2 would not be created for vowels, as this could lead to over-interpretation of unreliable data. Instead, mean vulnerability rates are shown in Table 6.5, while noting that (a) some of the findings are based on relatively few observations and (b) mean values may have been heavily influenced by one or two speakers. To aid interpretation, the table shows the number of occasions on which each vowel contrast was tested (assuming no missing data).

Vowel confusion (# occasions tested)	*Mean vulnerability rate*	*Vowel confusion (# occasions tested)*	*Mean vulnerability rate*
/i/ → /ɪ/ (2)	0.146	/ɪ/ → /i/ (5)	0.131
/ɔu/ → /œy/ (2)	0.104	/œy/ → /ɔu/ (1)	0.083
/i:/ → /y:/† (2)	0.188	/y:/ → /i:/† (1)	0.000
/ɛ/ → /ɪ/ (5)	0.185	/ɪ/ → /ɛ/ (3)	0.000
/u/ → /o:/ (3)	0.117	/o:/ → /u/ (2)	0.063
/a:/ → /ɑ/ (7)	0.089	/ɑ/ → /a:/ (4)	0.031
dip → mon (16)	0.081	mon → dip (9)	0.023
/ɛ/ → /ɑ/ (3)	0.076	/ɑ/ → /ɛ/ (6)	0.008
/i/ → /e:/ (5)	0.066	/e:/ → /i/ (3)	0.010
/ɪ/ → /ʏ/ (3)	0.040	Not tested	-
/ɔ/ → /o:/ (4)	0.021	/o: / → /ɔ/ (5)	0.017
/ɔ/ → /ɑ/ (7)	0.035	/ɑ/ → /ɔ/ (2)	0.000
/e:/ → /ɪ/ (3)	0.035	Not tested	-

† In Dutch phonology, this contrast only applies to words that end in /r/; the phonemes /i, y/ are lengthened to [i:, y:] in this context.

Table 6.5. Mean vowel vulnerability rates in the forced-choice study. The data are displayed in order of decreasing vulnerability (summed over the two directions) from top to bottom. The number of occasions on which each directional confusion was tested is shown in parentheses.

As can be seen, the number of occasions on which each confusion was tested is usually greater for one of the two error directions. This is because most vowel confusions were found to be strongly directional in the orthographic-transcription study, and there would have been little purpose in including distractors that are not representative of the types of error observed. For two of the vowel confusions, /i/ - /ɪ/ and /ɛ/ - /ɑ/, Table 6.5 reveals that the mean vulnerability rate was *higher* for the direction that was tested less often; this implies that the predominant error direction switched between the two response modes. It would not be prudent to read too much into this finding, however, due to the aforementioned issues regarding the reliability of central tendency measures. For example, the high mean error rate for /i/ → /ɪ/ was due to high error rates for just two speakers, while the remaining speakers did not yield *any* errors of this type. In contrast, /ɪ/ → /i/ confusions were observed in 5 out of 8 speakers. Accordingly, as shown below, when the vulnerability of each of these directional confusions is expressed in terms of the summed rank, the most error-prone direction is the same for both response modes (/ɪ/ → /i/).

As stated, the data in Table 6.5 should not be interpreted too closely. However, some general observations can be made. Firstly, two of the most vulnerable contrasts, /i/ - /ɪ/ and /ɛ/ → /ɪ/, were also prominent in the control population and may not be symptomatic of dysarthria (see Chapter 5, Fig. 5.4 and Table 5.6). The remaining confusions that were tested on a reasonable number of occasions (meaning that the findings can be considered reasonably reliable) yielded low error rates, at least in the majority of speakers. For example, the next most vulnerable category after /ɛ/ - /ɪ/, /u/ - /o:/, yielded just three errors in the direction /o:/ → /u/ across all speaker-listener observations. There were nine errors in the opposite direction, /u/ → /o:/, but six of these arose due to a single speaker.

The monophthong confusions are also depicted as theoretical movements across the vowel space (Fig. 6.6). In cases where the mean vulnerability rate in the non-dominant direction is at least one-third of the total mean vulnerability rate, two arrow heads are shown. The graph allows appreciation of the overall error pattern rather than trying to interpret results for specific confusions that, in general, have low error rates as well as low reliabilities. Monophthong confusions are also plotted for the free-response mode (Fig. 6.7), although the reader is reminded that the error metrics for the two response modes are conceptually different. Thus, a direct comparison across the two figures of the arrow thicknesses for a given vowel confusion is not recommended; rather, the purpose is to gain a visual impression of the most prominent errors in each mode. Note that Fig. 6.7 differs from the corresponding graph presented in Chapter 4 (Fig. 4.6) in several ways: (a) in the current graph, the thicknesses of the arrows are based on the raw data (i.e., the total number of

occasions on which the confusion was observed, summed over all speakers), rather than on the error metric devised in Chapter 4 (the MPE), as the raw data bear a closer relationship to MC vulnerability rates; (b) the error metrics were calculated using just the subset of participants that were assessed in the forced-choice study; and (c) data for confusions that were not tested in the forced-choice mode have not been displayed.

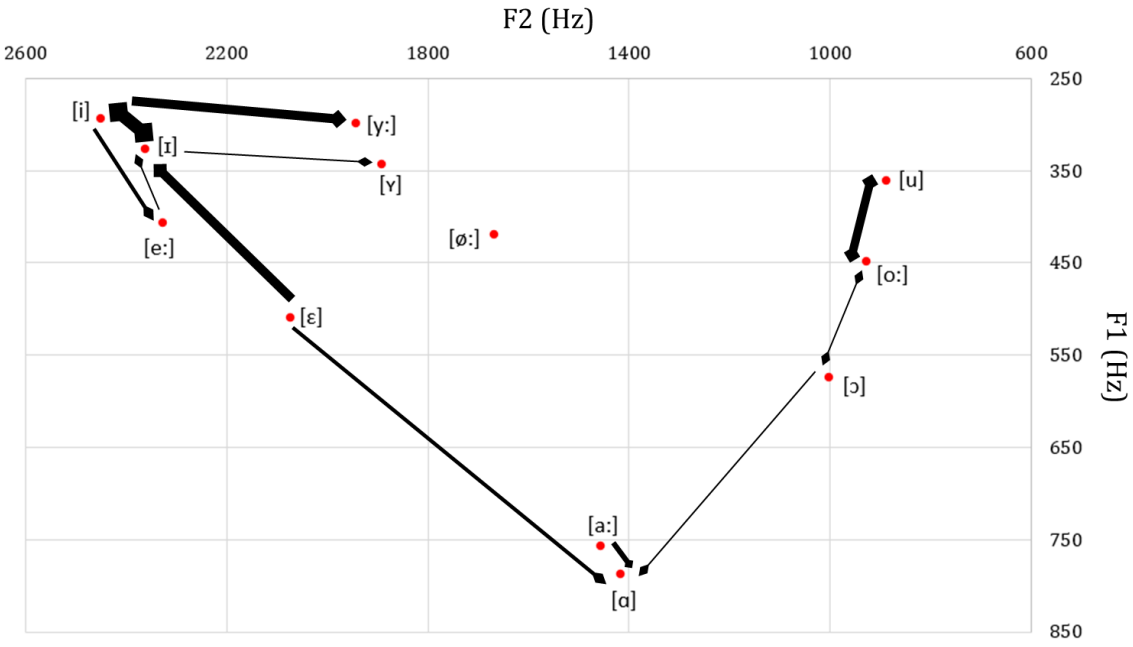


Figure 6.6. Monophthong confusions from Table 6.5 (forced choice). The thickness of each arrow is proportional to the mean vulnerability rate (summed over the two directions), while the arrow head indicates the predominant direction (two arrow heads denote bidirectional errors).

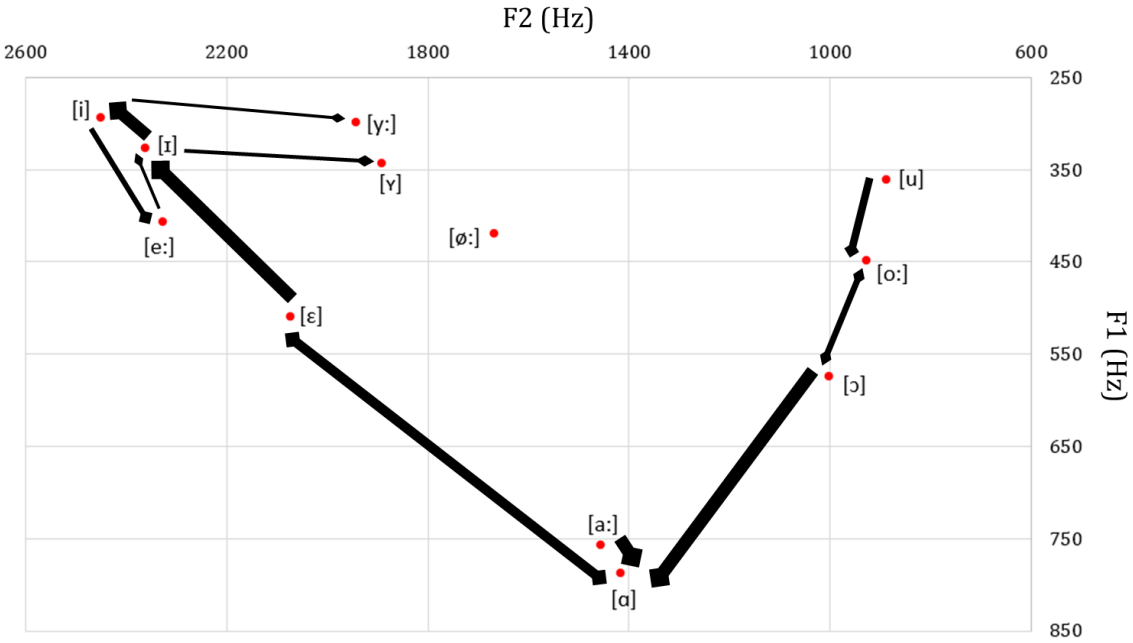
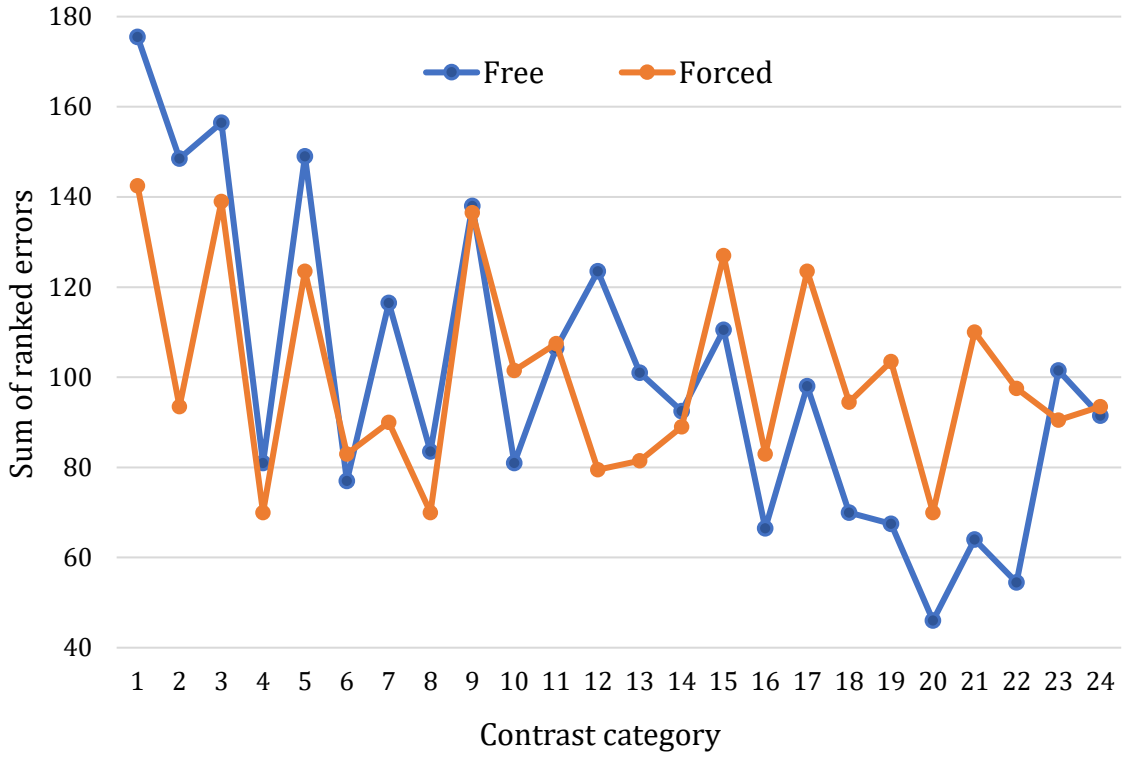


Figure 6.7. Monophthong confusions from the free-response study. The thickness of each arrow is proportional to the total number of occasions on which the confusion was observed, summed over both directions and over all eight speakers.

These graphs do not reveal any new insights; rather, they reinforce the findings mentioned above: the most prominent vowel confusions in the forced-choice mode (Fig. 6.6) are mainly confined to the top-left corner of the vowel space, as was also the case for neurotypical speakers. Meanwhile, the vowel confusions seen in the free-response study (Fig. 6.7) that involved large movements across the vowel space, /ɑ/ - /ɛ/ and /ɔ/ → /ɑ/, are no longer prominent. These errors were deemed to be "dysarthric" in Chapter 5, as they were found to be substantially lower (or even completely absent) in neurotypical speakers.

In summary, it seems that with the exception of vowel confusions that were also common in control speakers, vowel contrasts were reasonably robust in the forced-choice mode. It is possible that the difference between the two response modes for speakers with dysarthria can sometimes be explained in terms of functional load. For example, the monophthong-diphthong category pertains to vowel *classes*. Therefore, the free-response mode provides a very large number of opportunities for such confusions to arise. In contrast, in the forced-choice study, only monophthong-diphthong confusions that match the target-foil pair can be recorded. Yet that are other discrepancies that do not seem to have an obvious explanation. The most striking example is that in the free-response mode, there were 22 /ɑ/ → /ɛ/ confusions across all speaker-listener observations; yet there was only one such error in the forced-choice mode, despite the fact that this contrast was tested on multiple occasions using target words that had yielded errors in the free-response mode.

To conclude this analysis, correlation coefficients for the ranked errors in the two response modes are presented. These were calculated in the same manner as for consonant contrasts. The total ranks for the two response modes are presented in Figs. 6.8 and 6.9, while the correlation coefficients between the rankings for individual speakers are listed in Table 6.6. The reader is reminded that zero-error categories receive a higher ranking in the forced-choice mode than in the free-response mode (in fact, this difference is clearly illustrated in Fig. 6.8, as neither response mode yielded any errors for category 20). Thus, the most appropriate way of interpreting Fig. 6.8 is to examine the rank ordering of the vowel errors within each response mode and then to compare the two sets of findings. Accordingly, the top six vowel errors in the free-response mode, listed in order of decreasing vulnerability, are: (1) monophthongisation, (2) /ɛ/ → /ɪ/, (3) /a:/ → /ɑ/, (4) diphthongisation, (5) /ɪ/ → /i/ and (6) /ɑ/ → /ɛ/. The equivalent rankings for the forced-choice mode are: (1) monophthongisation, (2) /ɛ/ → /ɪ/, (3) /ɪ/ → /i/, (4) /i/ → /e:/, (5) /a:/ → /ɑ/ and (6) /u/ → /o:/, where the last two confusions yielded equal total ranks. Pearson's *r* for the correlation between the total ranked scores in the two response modes was 0.622, $p = 0.006$ (one-tailed).



1	monophthongisation	2	diphthongisation
3	/ɛ/ → /ɪ/	4	/ɪ/ → /ɛ/
5	/a:/ → /ɑ/	6	/ɑ/ → /a:/
7	/ɔ/ → /ɑ/	8	/ɑ/ → /ɔ/
9	/ɪ/ → /i/	10	/i/ → /ɪ/
11	/ɛ/ → /ɑ/	12	/ɑ/ → /ɛ/
13	/ɔ / → /o:/	14	/o:/ → /ɔ/
15	/i/ → /e:/	16	/e:/ → /i/
17	/u/ → /o:/	18	/o:/ → /u/
19	/i:/ → /y:/	20	/y:/ → /i:/
21	/ɔu/ → /œy/	22	/œy/ → /ɔu/
23	/e:/ → /ɪ/		
24	/ɪ/ → /ʏ/		

Figure 6.8. Sum of ranked errors for the directional vowel confusions as judged via the forced- and free-response modes. The confusion codes are shown in the legend beneath the figure. The last two confusions were only tested in one direction in the multiple-choice mode.

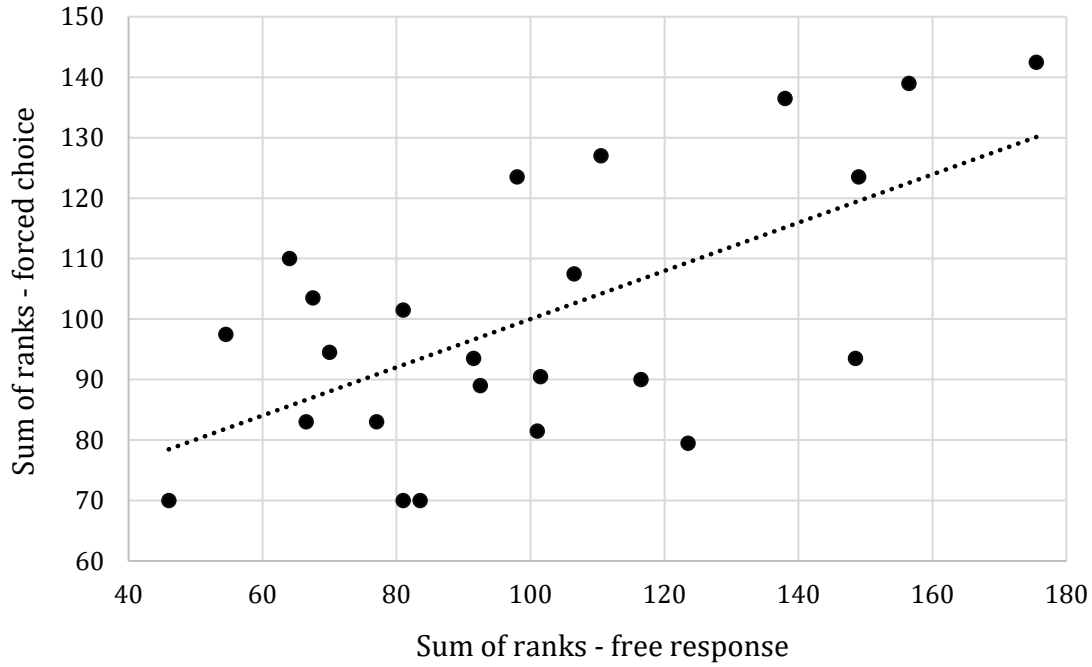


Figure 6.9. Relationship between the total sum of the ranks in the two response modes for the 24 vowel categories. Pearson's *r* = 0.62, *p* = 0.006 (one-tailed).

Speaker ID	*Pearson's r*	*p (one-tailed)*
S7	0.189	0.189
S6	0.399	0.027
S2	0.861	0.000*
S3	0.455	0.013
S5	0.338	0.053
S8	0.172	0.211
S9	0.621	0.001*
S4	0.753	0.000*

Table 6.6. Pearson's *r* for the vowel error rankings in the two response modes for each speaker, where the speakers are displayed in order of decreasing intelligibility (based on the free-response study). The *p*-values marked with an asterisk are significant assuming a Bonferroni-corrected alpha level of (0.05 / 8) = 0.0063.

As expected, the correlation between the two response modes is weaker for vowel contrasts than for consonant contrasts, due to the fact that the former involve specific phoneme pairs. This is particularly noticeable for individual speakers, where Pearson's *r* is only significant for three speakers (as opposed to seven in the case of consonant contrasts).

6.4. Discussion

6.4.1. Word-accuracy scores and speaker intelligibility rankings

Based on the literature review, it had been hypothesised that word-accuracy scores (the percentage of correct words) would be higher in the forced-choice mode than in free response. This was found to be the case for all eight speakers, with a mean absolute difference (± 1 SD) of 13.1% ± 6.9%. The increase in accuracy was greater for speakers of lower intelligibility, which is likely to be largely due to the fact that there is a ceiling effect for speakers of higher intelligibility. Bunton and Weismer (2001) reported some preliminary findings in which the forced-choice format actually encouraged errors relative to an open response mode. The implication, presumably, is that when a token is produced in a distorted manner, the act of presenting the listener with the non-target option can make them question what they heard and start to 'look for' (and sometimes find) the error. It will be shown in Section 6.4.2 that this phenomenon also occurred in the present study. In other words, errors were reported in the MC mode for tokens that had been transcribed 100% correctly in the free-response paradigm. However, given the universal increase in word accuracy scores across speakers, it is clear that the opposite scenario was far more common – a large number of errors "disappeared" in the forced-response mode, even if they had been perceived by all listeners in the free-response study.[5] This suggests that errors that are distortions rather than substitutions are more likely to be perceived as the intended target when the listener's options are constrained than when they are unconstrained. The underlying reasons for this are discussed in Section 6.4.2.

A further observation was that the speakers were not ranked in precisely the same order of intelligibility in the two response modes, in common with the findings of Vigouroux and Miller (2007) for speakers with Parkinson's disease. Nevertheless, there was a reasonably strong correlation between the two word accuracy scores (r = 0.86, one-tailed p = 0.003), higher than that reported by Vigouroux and Miller for their dysarthric group (r = 0.72, two-tailed p < 0.001). Due to the low sample size in the present study, it would be unwise to attempt to interpret this discrepancy, but as discussed in Chapter 2 (and as also noted by Vigouroux and Miller), one of the most important factors affecting the degree of correlation is likely to be the range of intelligibility levels among subjects, with stronger correlation expected in populations where there are larger gaps between abilities. Furthermore, due to the ceiling effect, the correlation is likely to be lower for populations in which the average

[5] Note that this statement is not referring to errors that *could* not be perceived in the MC study, due to the fact that the error in question was not included as one of the distractors. This situation arose as well, but inspection of the data showed that it was not the main cause of the increase in word accuracy between the two response modes.

intelligibility is higher. A third factor to consider, and perhaps the most important, is the way in which the distractors were chosen. In the present study, they were based on the findings of orthographic transcription of the same speech data, which would naturally increase the degree of correlation relative to studies where this was not the case. As for the clinical implications of different rankings in the two response modes, the question naturally arises: Which of the two accuracy values is a truer reflection of the speaker's level of functional impairment? This could be addressed in future research by examining the degree of correlation between each type of word-accuracy score (open and closed) and an intelligibility measure derived from spontaneous speech.

6.4.2. Similarity of error profiles in the free- and forced-response modes

Before discussing the findings with regard to the similarity in error profiles, it is worth reminding the reader that the level of inter-rater agreement in this study was not high. The values obtained for Fleiss' kappa represented "moderate agreement" for six out of eight speakers, with "fair agreement" and "slight agreement" for the remaining two speakers respectively. As discussed in the free-response study (see Chapter 4, Section 4.4.4), there are many reasons for supposing that the level of inter-rater agreement could be improved, including the fact that stronger agreement would be expected when the metric pertains to the final outcome measure (the error profile) rather than to individual test items. This is because different listeners might perceive the same contrast error on different targets. Nevertheless, the reader should bear in mind that the findings reported in this study may have low to moderate levels of reliability and validity. Therefore, the following discussion is mainly limited to trends observed across the entire cohort and for multiple contrast categories. Findings of a more specific nature (e.g., for an individual speaker) are only mentioned by way of an illustrative example or when, to the best of the author's judgment, it is highly unlikely that the result is peculiar to the current set of listeners.

In the case of both vowel and consonant contrasts, there were differences in the top six error categories (defined on the basis of summed ranks) identified by the two response modes. At the level of the individual, the two sets of error ranks generally showed poor to moderate agreement, as evidenced by the Pearson's *r* values in Tables 6.4 and 6.6. The correlation values were lower for vowels than for consonants, which is unsurprising given that the vowel categories were defined in a much narrower way (i.e., as substitutions between two specific phonemes). In order to increase the level of agreement between the two techniques in the case of vowel errors, it is likely to be necessary to devise a method of assigning vowel substitutions to broader categories.

It was argued that some of the differences between the two response modes could be explained by the fact that functional load acts a confounding factor in the free-response mode but not in the forced-choice paradigm. Such differences are unlikely to present a problem in the long term. If it proves to be the case that the free-response mode is more appropriate for clinical practice, then acquisition of data from a large number of speakers with different dysarthria severities and types, as well as from a large control population, would make it possible to define threshold error levels above which a particular contrast category can be considered problematic for the individual in question. However, there were also error categories that almost "disappeared" in the forced-choice mode, despite the facts that (a) there were ample opportunities for the error to be perceived based on the list of distractors and (b) the confusion did not seem to have an exceptionally high functional load in the free-response mode. Examples included final consonant addition, /ɑ/ → /ɛ/ and /ɔ/ → /ɑ/. Furthermore, if one examines the data of specific speakers (rather than the cohort as a whole), it can be seen that certain individuals "lose" a significant number of error categories in the forced-response mode. For example, S5 yielded 8 instances of voicing (of phonologically voiceless consonants) in the free-response study, whereas this category was 100% robust in the MC mode. In the 'nasal place' category, he yielded 12 fronting errors and 2 backing errors in orthographic transcription. This was reduced to just one error in each direction in the forced-response mode. For two of the words that had yielded a nasal place error in the free-response study, all listeners had transcribed the error in the same way. There are likely to be numerous underlying causes for disappearing errors, the most obvious of which are described in the following paragraphs. Some of these mechanisms are mutually exclusive, while others may operate in parallel.

(1) It was sometimes the case that the error perceived by listeners in the free-response study was not included among the distractors. There were three situations in which this arose. Firstly, there were some errors that were not sufficiently common to be tested in the MC study and/or did not meet Kent et al.'s (1989) criterion of a contrast in a single phonetic feature (e.g., /l/ vs. /n/ confusions). Therefore, speakers who tend to yield large numbers of atypical errors may achieve an artificially high accuracy in an MC paradigm (assuming that the atypical errors are not always coded as another type of error), which would be a distinct disadvantage of the approach. Secondly, there were occasions where the error transcribed in the free-response study *did* meet the Kent criterion, but it was not chosen as one of the distractors for that particular word. Thirdly, there were phonetic-contrast categories that were too broad to include all possible manifestations of the error perceived in the cohort. For example, there are numerous consonant phonemes that could be appended to the end of the word /du/ ('*(I) do*') to produce a meaningful word of Dutch.

However, if one wishes to test the vulnerability of final consonant addition for this target in the multiple-choice study, it is only possible to include a maximum of three of these options as distractors (and in practice, the number tended to be lower, to enable other phonetic-contrast errors to be tested for the given word). This might explain why the category 'final consonant vs. null' disappeared in the MC mode. In fact, one could argue that all of the syllable-shape categories (with the exception of /h/ vs. null) are disadvantaged by the MC mode for this reason, and that it is not logical to include them alongside categories that are phoneme-specific.

(2) The second main cause of disappearing errors is that a large proportion of the misarticulations of speakers with dysarthria are *distortions*. It seems logical that, under certain circumstances, such productions would be more likely to be scored as the target when the listener uses a closed (as opposed to an open) response mode. For example, if the target has a lower lexical frequency than the potential substitution, it might be less likely to be considered by the listener in a free-response mode. However, once the listener is presented with both options, these have an equal chance of being chosen. There were cases in the data where it was likely that this was the underlying explanation. For example, for Speaker 4, the token /sɔp/ (English *sud*, as in 'soap sud') was often transcribed orthographically as /sɑp/ (*juice*), a word with much higher lexical frequency. This error disappeared in the forced-choice mode, suggesting that although the phoneme /ɔ/ was distorted *towards* /ɑ/, it in fact bore closer resemblance to the intended target. A further reason why certain distortions are prone to disappearing is that they may have to compete with other distortions that are tested simultaneously. In other words, since the listener can only choose one of the distractors, if a word is produced with multiple distorted phonemes, the error that is most "prominent" will dominate, where the concept of prominence incorporates both articulatory considerations (the extent to which the sound is distorted) and perceptual considerations (the perceptual similarity of the two phonemes). As an example, consider Speaker 4's realisation of the word /krɔm/ (*crooked*). This was transcribed as either /krɑp/ (*narrow*) or /klɑp/ (*clap*, *smack*) in the free-response study, showing that errors were perceived at all three word positions. In the MC study, the target was coded as either /klɔm/ (*climbed*) or /krɔp/ (*head*, as in 'head of lettuce'), while the vowel distractor /krɑm/ (*clamp*) was not chosen. Thus the consonant errors predominated and effectively masked the vowel error. It is possible that this particular finding (a preference towards consonant errors) holds more generally, as listeners tolerate greater phonetic variation in vowels than in consonants before they perceive a different phoneme (Haley et al., 2000). It might explain why the only *consonant* category to disappear completely was 'final consonant vs. null', which, as mentioned above, is likely to have

disappeared because it was not phoneme-specific. For vowels, on the other hand, the /ɑ/ - /ɛ/ and /ɔ/ - /ɑ/ confusions showed very low vulnerability rates, despite the fact that they are categories involving specific phonemes.

(3) Finally, it is possible that the multiple-choice approach may *bias* the listener towards certain responses. This was already mentioned in the previous paragraph, where it was pointed out that items that simultaneously test vowel contrasts and consonant contrasts might confer an "unfair" advantage on consonant contrasts (in the sense that if a vowel and a consonant have the same degree of articulatory distortion, the consonant error might be more likely to be chosen). Similarly, items that pit a pre-vocalic contrast against a post-vocalic contrast could be considered biased because initial consonants are more easily identifiable than final consonants (Redford & Diehl, 1999). Furthermore, an astute observer could use the strategy of "eliminating outliers" (Poundstone, 2015: Chapter 3) to increase his/her chances of choosing the intended target. According to this strategy, choices that are incongruent with the others in an MC question may be dismissed, as they have a higher likelihood of being incorrect.[6] Thus, in the test item <u>*boon*</u> – *bon boom boen*, the middle distractor (*boom*) stands out as being different, even if one has no knowledge of phonetics or of the goals of the investigation.[7] The author was aware of all of these potential sources of bias and tried to minimise their occurrence by including as many test items as possible that (a) did not pit vowel contrasts against consonant contrasts or C1 contrasts against C2 contrasts and (b) did not contain an obvious outlier. An example of such an item was <u>*taal*</u> – *daal paal kaal*. However, with the numerous constraints acting on the design of the MC assessment, it was rarely possible to achieve items that were free of the aforementioned sources of bias. The potential for bias in the multiple-choice version of Kent et al.'s (1989) test has been noted previously (Bunton et al., 2007). These authors, who investigated English-speaking adults with Down syndrome, also observed differences between the free- and forced-response modes. However, these differences did not match those reported in the present study. In particular, they found that listeners were biased towards *vowel* errors in the MC mode when an error existed on both the vowel and a consonant phoneme simultaneously. The discrepancies between the two studies may be due to differences in aetiology and language. Furthermore, the broad transcription in

[6] A simple example would be: "What is the square root of 64: (a) 7, (b) 8, (c) 9 or (d) 32?"

[7] To the best of the author's knowledge, it has not been investigated whether listeners might employ such a strategy in a multiple-choice perception test carried out for research purposes (where there is little to be gained from choosing the "right" answer). However, the existence of the phenomenon seems plausible, especially if (a) one includes the possibility that it takes place at a subconscious level and (b) the listeners have some knowledge of the research field (e.g., SLTs and phoneticians).

Bunton et al.'s (2007) study was carried out by experts, while the multiple-choice responses were provided by lay listeners. Future cross-linguistic research on the interaction between a speaker's misarticulations and the listener's response paradigm would be worthwhile. In addition, further development of the dysarthria assessment proposed in this thesis could result in improvements with regard to issues such as bias. A relatively simple modification would be to increase the number of distractors for each target. This would reduce the potential for listeners to use the strategy of eliminating outliers. It would also mitigate against another problem raised in the above discussion, namely that it was not always possible to include every error perceived in the free-response study among the set of distractors in the multiple-choice study.

As mentioned in Section 6.4.1, there were also instances of errors being *enhanced* in the forced-response mode, or of new errors appearing, although these two processes were far less common than error reductions. An example of enhancement was /tɑk/ (*branch*) transcribed as /dɑk/ (*roof*) by one listener in the free-response mode, but by all three listeners in the forced-choice mode. An example of a new error was /rɛi/ (*row*, *queue*) → /vrɛi/ (*free*), which was chosen by two out of three listeners in the MC mode while in orthographic transcription, the target word had yielded no errors of any kind. It is possible to imagine at least two causes of new or enhanced errors: (1) the substitution has lower lexical frequency than the target, meaning that it was less likely to be considered by the listener in the free-response mode, and (2) the aforementioned psychological phenomenon whereby offering the error as an option causes the listener to look for (and find) evidence of its existence.

6.4.3. Important phonetic-contrast errors in Belgian Dutch dysarthria

It was stated in Chapter 4 that phonetic-contrast errors are most appropriately discussed in the light of the findings of the normal-control and multiple-choice studies. That is, by integrating the data from Chapters 4 and 5 with the error profiles obtained in the present study, it becomes possible to gain a preliminary idea of which contrast categories are likely to be "important" in the current cohort. When designing the thesis, it had been expected that the three features listed below would be used to classify a contrast error as important. However, as shown in the commentary beneath each criterion, the current data did not always provide sufficient evidence to make a definitive judgement. Therefore, the following analysis should be considered preliminary in nature and in need of corroboration in future studies. The three criteria were defined as follows:

1) An "important" error should be **prominent in the free-response mode**, even if this is only for one speaker. This is because if the error does not arise frequently in an unconstrained response mode, using a phonemically-balanced word list, then it is unlikely to have a substantial effect on real-world intelligibility. There was no clear guideline for defining an error as "prominent". However, from inspecting the raw data, it was decided that a prominent *consonant* contrast-error would be one that was observed on at least 6 occasions in at least one of the ten dysarthric speakers. This would mean that, for speakers who were only assessed by three listeners, the threshold would equate to a situation where at least two separate tokens were perceived as the directional contrast by all listeners (although note that a total of 6 errors could also arise in other ways, e.g., two errors perceived for each of three different targets). In the case of *vowel* contrasts, it was deemed that a threshold of three errors in at least one speaker would be more reasonable. This lower threshold was chosen because firstly, there is only one vowel phoneme per word, as opposed to two consonant phonemes, and secondly, the vowel categories are more specific than the consonant categories, such that the *a priori* probability of an error in any given category is lower.

2) The confusion should be **dysarthric**, meaning that it yields a significantly greater number of errors in speakers with dysarthria than in control speakers. It was only possible to test this property formally for some of the contrast categories (see Chapter 5, Table 5.5). For other categories, there were too few errors in one or both populations to conduct a statistical test. In the following summary, therefore, an error is assumed to be dysarthric unless there was evidence to the contrary in Table 5.5. Further research may reveal that some of the errors currently labelled as dysarthric are in fact equally common in neurotypical speakers from Antwerp (e.g., initial /h/ deletion) or sufficiently common such that a cut-off error rate will need to be established in order to consider the contrast to be problematic.

3) The category should be **vulnerable** in the sense that it yields errors with some degree of consistency in the multiple-choice mode. The purpose of this criterion was to rule out errors that were only prominent in the free-response mode because the contrast in question had an exceptionally high functional load. In other words, there would be little purpose in delivering intervention for a contrast that is robust in the sense that it is realised correctly on the vast majority of occasions. In previous research that used Kent et al.'s (1989) approach with forced-choice responses (e.g., Kent et al., 1990; Whitehill & Ciocca, 2000b), the mean error proportions across the dysarthric speakers were not

particularly high: typically < 0.2 for most contrast categories.[8] In subgroups with mild dysarthria, the mean error proportions were even lower, although the top few categories yielded error rates of the order of 0.1-0.2. Therefore, in the current analysis, a directional contrast was considered "vulnerable" if it yielded an error proportion of at least 0.1 in at least one of the eight speakers assessed in the forced-choice study. Note that no judgment about vulnerability was made for the 'initial consonant vs. null' and 'final consonant vs. null' categories, which yielded no errors in the MC study, nor for 'monophthong vs. diphthong' confusions. As discussed in Section 6.4.2, these categories may have been disadvantaged due to the fact that it was not possible to test all possible manifestations of the error. Note further that a vulnerability threshold of 0.1 will deem categories such as /ɑ/ → /ɛ/ to be unimportant, despite the fact that such categories *may* have disappeared in the forced-choice mode for spurious reasons (see Section 6.4.2). Therefore, the findings in relation to the third criterion should be viewed as preliminary, as further development of the dysarthria assessment may reduce such sources of bias and render these categories more vulnerable.

The findings for the above three criteria are presented in Table 6.7 (for consonants) and Table 6.8 (for vowels). These tables summarise the body of evidence acquired in this thesis regarding the importance of each phonetic-contrast category. The categories have been numbered in the same manner as in Figs. 6.3 and 6.8, to facilitate comparison with the ranking data. Directional errors that are highlighted in grey are classified as important in Dutch speakers with dysarthria based on the current findings. The reader is cautioned that the judgments made for many of the vowel categories (Table 6.8) were based on data of low reliability, owing to the fact that vowel confusions were not consolidated into categories. Therefore, with the exception of a few categories that pertained to very common vowel phonemes, the number of potential errors was low. In general, however, the approach taken in this analysis was to set the thresholds for labelling an error as "prominent" or "vulnerable" at fairly low levels. Likewise, errors were considered "dysarthric" unless there was strong evidence to the contrary. This cautious approach was adopted since it errs on the side of ensuring that any errors that *might* be important are tested in future dysarthria assessments. The alternative would have been to risk dismissing errors that are in fact important, but could not be detected in the present thesis due to low power.

[8] Note, however, that unlike the present study, these publications did not calculate separate error rates for the two directions (e.g., voicing and devoicing), meaning that the error proportion for the predominant direction in each category would have been higher.

Category ID	*Directional consonant contrast*	*Prominent in free mode?*	*Dysarthric?*	*Vulnerable in forced mode?*
1	Devoicing (stops, fricatives)	Yes	Yes	Yes
2	Voicing (stops, fricatives)	Yes	No	Yes
3	Stop backing	Yes	Yes	No
4	Stop fronting	No	Yes	No
5	Nasal backing	Yes	No	Yes
6	Nasal fronting	Yes	No	Yes
7	Fricative backing	Yes	Yes	Yes
8	Fricative fronting	No	Yes	No
9	Stop → fricative	Yes	Yes	Yes
10	Fricative → stop	Yes	Yes	Yes
11	Nasal → stop	Yes	Yes	Yes
12	Stop → nasal	No	Yes	No
13	/r/ → fricative	Yes	Yes	Yes
14	fricative → /r/	Yes	Yes	Yes
15	/r/ → /l/	Yes	Yes	Yes
16	/l/ → /r/	Yes	Yes	Yes
17	Initial /h/ addition	No	Yes	Yes
18	Initial /h/ deletion	Yes	Yes	Yes
19	Null → initial consonant	No	Yes	-
20	Initial consonant → null	No	Yes	-
21	Null → final consonant	Yes	Yes	-
22	Final consonant → null	No	Yes	-
23	Initial singleton → cluster	Yes	Yes	Yes
24	Initial cluster → singleton	Yes	Yes	Yes
25	Final singleton → cluster	Yes	Yes	Yes
26	Final cluster → singleton	Yes	Yes	Yes

Table 6.7. Importance of consonant contrast categories in Dutch dysarthria, as judged by three criteria. The categories highlighted in grey are classed as "important". The numbers in Column 1 facilitate comparison with Fig. 6.3.

Category ID	*Directional vowel contrast*	*Prominent in free mode?*	*Dysarthric?*	*Vulnerable in forced mode?*
1	Monophthongisation	Yes	Yes	-
2	Diphthongisation	Yes	Yes	-
3	/ɛ/ → /ɪ/	Yes	No	Yes
4	/ɪ/ → /ɛ/	Yes	Yes	No
5	/a:/ → /ɑ/	Yes	Yes	Yes
6	/ɑ/ → /a:/	Yes	Yes	Yes
7	/ɔ/ → /ɑ/	Yes	Yes	Yes
8	/ɑ/ → /ɔ/	Yes	Yes	No
9	/ɪ/ → /i/	Yes	No	Yes
10	/i/ → /ɪ/	Yes	Yes	Yes
11	/ɛ/ → /ɑ/	Yes	Yes	Yes
12	/ɑ/ → /ɛ/	Yes	Yes	No
13	/ɔ/ → /o:/	Yes	Yes	Yes
14	/o:/ → /ɔ/	Yes	Yes	No
15	/i/ → /e:/	Yes	Yes	Yes
16	/e:/ → /i/	No	Yes	No
17	/u/ → /o:/	Yes	Yes	Yes
18	/o:/ → /u/	Yes	Yes	Yes
19	/i:/ → /y:/	Yes	Yes	Yes
20	/y:/ → /i:/	No	Yes	No
21	/ɔu/ → /œy/	Yes	Yes	Yes
22	/œy/ → /ɔu/	No	Yes	Yes
23	/e:/ → /ɪ/	Yes	Yes	Yes
24	/ɪ/ → /ʏ/	Yes	Yes	Yes

Table 6.8. Importance of vowel contrast categories in Dutch dysarthria, as judged by three criteria. The categories highlighted in grey are classed as "important". The numbers in Column 1 facilitate comparison with Fig. 6.8.

The implications of the findings in Tables 6.7 and 6.8 with regard to enhancing our understanding of speech production in dysarthria are discussed in Chapter 8. From a methodological perspective, the findings could be useful for refining the dysarthria assessment proposed in this study, or for developing other Belgian Dutch dysarthria assessments, as they suggest that certain contrast categories may not need to be tested. However, such findings would need to be confirmed in other studies with larger sample sizes and a wider range of dysarthria types. As mentioned, these studies should pay particular attention to vowel categories that were prominent and dysarthric, but not vulnerable, as low vulnerability rates in the MC study could be due to bias. However, the effect of bias cannot simply be assumed to be undesirable; when an error disappears in the forced-choice mode, this could imply that it may not have a significant impact on a speaker's intelligibility in everyday speech.

6.5. Summary

The goal of this study was to determine the differential effect of the open and closed response formats on word-accuracy scores and phonetic-contrast error profiles in speakers with dysarthria. The percentage of correct words was found to be significantly higher in the forced-response mode, with a mean absolute difference (± 1 SD) of 13.1% ± 6.9%. It was surmised that this difference can be partly attributed to the fact that, relative to the closed mode, the open mode is more likely to cause a phonetic distortion to be perceived as a phonemic substitution. However, it is also possible that some genuine substitution errors go undetected in the forced-choice mode due to bias. The speakers were not ranked in precisely the same order in the two modes; however, there was high correlation between the two sets of word-accuracy scores (r = 0.86, one-tailed p = 0.003). Fleiss' kappa was used to determine the level of agreement on responses to individual test items. There was moderate agreement for six out of eight speakers, with fair agreement and slight agreement for the remaining two speakers. Further work is needed to determine inter-rater (as well as intra-rater) reliability for the actual outcome measure, i.e., the profile of phonetic-contrast errors. For both vowels and consonants, there were differences in the top six error categories identified by the two response modes. For consonant contrasts, the correlation between the summed ranked errors for the two response modes was r = 0.735 (one-tailed p < 0.001). The corresponding value for vowels was 0.622 (p < 0.01). Lower correlation was observed for individual speakers, particularly in the case of vowels. Overall, these findings imply that the two response modes provide qualitatively different information. Finally, the results of the current study were integrated with the findings of

Chapters 4 and 5 to obtain a preliminary indication of the vowel and contrast confusions that are important in Belgian Dutch dysarthria.

The next chapter presents the final investigation of the thesis, which is a departure from the previous investigations in the sense that it is not concerned with the methodology of identifying and categorising segmental, articulatory errors. Rather, it addresses the underlying premise for the clinical value of conducting articulatory analysis, namely that the errors identified by such techniques are detrimental to spontaneous-speech intelligibility. Ideally, this question would be addressed using an explanatory approach. However, this would be a significant undertaking that was beyond the scope of the present thesis. Therefore, the goal was to determine the degree of correlation between measures of intelligibility derived from single-word reading in speakers with dysarthria (as reported in the current chapter and in Chapter 4) and an intelligibility metric derived from spontaneous speech. A moderate to high correlation would be a necessary, but not a sufficient condition for arguing that articulatory therapy is likely to lead to a worthwhile improvement in intelligibility in spontaneous speech.

7. Study 4: Correlation between single-word intelligibility and spontaneous-speech intelligibility in speakers with dysarthria

7.1. Objectives

The final investigation reported in this thesis addresses the underlying assumption of the previous studies in Chapters 4-6, namely that errors identified by phonetic-contrast analysis are detrimental to real-world intelligibility. As discussed in Chapter 2, a thorough investigation of this question requires an explanatory approach and would have been beyond the scope of the project. Therefore, it was decided that an investigation would be carried out, for the speakers with dysarthria only, which had the limited objective of examining the degree of correlation between intelligibility measures derived from single-word reading (SWR) and an intelligibility measure derived from spontaneous speech. A moderate to high correlation coefficient would indicate, at the very least, that substitution errors perceived in single-word reading co-vary with the factors that affect intelligibility in spontaneous speech. If a strong correlation is *not* observed, then future research could focus on identifying the subset of speakers for whom articulatory therapy is expected to be beneficial. The first objective of the present study was to answer the following question:

What is the level of correlation between metrics of intelligibility derived from single-word reading and a metric derived from spontaneous speech?

The spontaneous-speech intelligibility (SSI) metric employed in the present study, which was based on the metric proposed by Lagerberg et al. (2014), had not previously been tested in speakers with dysarthria. It was chosen following a thorough literature review, which led to the conclusion that a technique based on *orthographic transcription* would be most appropriate, especially given the characteristics of the current listening population. Transcription requires listeners to report the perceived speech output, a task that is expected to be more objective, and to require less skill and experience, than that of providing an intelligibility rating. Given that the Lagerberg metric had not previously been tested in speakers with dysarthria, the second objective of Study 4 was as follows:

To assess the suitability of the Lagerberg et al. (2014) metric for quantifying spontaneous-speech intelligibility in speakers with dysarthria.

The second objective was exploratory; i.e., no specific hypothesis or question was posed. Although the study was limited to assessing one particular SSI metric, the findings were expected to have broader relevance for understanding the challenges associated with the quantification of spontaneous-speech intelligibility, especially in cases where a transcript of the speech sample is not available.

7.2. Method

7.2.1. Calculation of spontaneous-speech intelligibility

Lagerberg et al.'s (2014) approach was described in Chapter 3 (Section 3.5.2). However, to briefly remind the reader of the general principles: monologues are divided into utterances that, as far as possible, coincide with the semantically natural pauses produced by the speaker. Listeners are instructed to transcribe each utterance in turn, using orthography to represent every word that they can understand and denoting every syllable perceived in the remaining (unintelligible) portions of speech with the symbol '0'. A word should be considered to be "understood" if the listener has a reasonable level of certainty that it corresponds to the intended target, even if the word was produced with distortion. However, if the listener is *not* reasonably certain of the intended target, they should not guess. For a given monologue and listener, the metric of spontaneous-speech intelligibility is the number of syllables in the words that are *intelligible* (and hence orthographically transcribed) divided by the *total* number of syllables perceived by the listener.

Several modifications were made to Lagerberg et al.'s (2014) approach. Firstly, listeners were provided with a title for each monologue so that they were aware of the general context, as would generally be the case in a real-world communicative situation where a speaker is relating a personal narrative. Secondly, the methodology for calculating the intelligibility metric was changed in two ways: (1) the intelligible words, which contribute to the numerator of the metric, were identified based on a *consensus* approach rather than relying on the judgments of individual listeners and (2) the number of syllables in the non-intelligible words (i.e., any words that had not been agreed upon using the consensus method) was determined solely by the author; thus the syllable counts of the listeners were discarded and did not contribute to calculating the denominator. These two changes are now described in greater detail, along with an explanation as to why they were needed.

As explained in Chapter 3 (Section 3.5.2), due to time constraints and the fact that many of the listening sessions were performed online, it was not possible to carry out the rigorous training method employed by Lagerberg et al. (2014). Consequently, a fairly common occurrence was that listeners did not seem to heed the instruction to not guess. Evidence of guesswork included large inter-listener differences in the transcribed utterances as well as transcriptions that were semantically implausible, e.g., *Ik ben dochter en de schoonzon* ("I am daughter and the son-in-law"). A further tendency among some listeners was to transcribe a stream of words that was almost devoid of meaning, e.g., "*Maar als ik dan dan zegt hij de 00*" ("But if I then then he says the 00"). It seems highly unlikely that these exact words were clearly perceived; yet they do not hold enough meaning to have been implied

by the context. Therefore, it appears that listeners who produced such transcriptions were using monosyllabic words (especially function words) as a means of approximating the unintelligible phonetic output of the speaker, perhaps because this was easier or quicker than counting syllables. It seems likely that if an extended training session had been administered, including the provision of specific feedback to listeners on their transcription efforts, instances of guesswork would have been substantially reduced.[1] Therefore, these suboptimal transcriptions do not necessarily imply that the method is unsuitable for future use in speakers with dysarthria. As for the present study, a correction technique was devised and implemented to deal with guesswork, one that was judged to result in a reliable estimate of the numerator of the SSI metric.

The goal of the adjustment was to disregard orthographic transcriptions that appeared to be guesswork. This was achieved using a consensus approach (see Table 7.1), the essence of which can be described as follows. For each word transcribed by each listener, the transcription was compared with those of the other listeners for the same (approximate) position in the utterance. In cases where the transcribed word was perceived in the same way by the majority ($\geq 50\%$) of the listeners,[2] the listener's transcription was considered as accurate (blue font in Table 7.1). Otherwise, the transcribed word was disregarded and did not contribute to the numerator of the SSI metric (the number of intelligible syllables) for the listener in question. For the example utterance shown in the table, it turned out that all listeners yielded the same number of intelligible syllables. However, this was not usually the case, as listeners tended to vary in terms of the number of words they had transcribed that met the criterion for consensus with other listeners.

In fact, the procedure was somewhat more intricate than that summarised above, as multisyllabic words were analysed on a syllable-by-syllable basis. For example, if the majority of listeners had transcribed a past participle of the form *ge-* [+stem] *-en* (e.g., *gesproken* – "spoken") at a given utterance position, then the inflectional morphemes *ge-* and *-en* were counted as intelligible syllables, irrespective of the inter-listener variation in the transcribed verb stem. This approach essentially rewarded the speaker for having provided enough information to signify that a particular inflectional form (e.g., a past participle) was uttered. Inflectional morphemes that served a different *function* in the two

[1] This statement is based on the fact that Lagerberg et al. (2014) did not report the same problem. However, their listeners were SLT students or graduates, which may have resulted in greater success in implementing the technique. Differences in the clinical population could have also played a role.

[2] There were between 3 and 5 listeners per speaker.

words, on the other hand, were not treated as a match. An example of this can be seen in Table 7.1. Two of the five listeners perceived the word *gewonnen* (the past participle of "to win"), which, in the author's opinion was indeed the word attempted by the speaker (see table caption). A third listener perceived the word *gewoon*, which is phonetically similar to *gewonnen*, but is an adjective meaning "common" or "usual". The role of the morpheme *ge-* is different in these two cases – a prefix denoting a past participle in the case of *gewonnen* and an adjectival morpheme used before a verb in the case of *gewoon*.[3] Therefore, it was reasoned that the speaker had not provided enough information for the majority of listeners to recognise the word class, and that the "*ge*" syllable would be scored as unintelligible. It could have been argued that since this was a test of intelligibility and not comprehensibility, all syllables that were transcribed the same way by the majority of listeners should be counted as accurate, irrespective of their meaning within the utterance. However, the chosen method was deemed to be more consistent with the notion that listeners should be "reasonably certain" of what they heard.

Listener	*Transcription*	*# intelligible syllables*
1	000 gewonnen heeft of verloren	4
2	want dan gaan we 00 00 of verloren	4
3	of dat hij gewonnen heeft of verloren	4
4	000 gaan we niet of verloren	4
5	of wat heeft gewoon niet of verloren	4

Table 7.1. Technique used to correct for guesswork. The speaker was describing his grandson's football match, and in the author's estimation, the best possible transcription of the utterance would have been *00 gewonnen heeft of verloren* ("00 has won or lost"), where the zeroes denote unintelligible syllables. For each listener, the transcribed words that were classed as "intelligible" based on the consensus method are shown in blue font.

The second modification pertained to the denominator of the intelligibility measure: the total number of perceived syllables. It was decided that this number would be *fixed* for a given utterance by a given speaker. In other words, the syllable counts carried out by the listeners were disregarded, and an estimate derived by the author was used in its place. For the utterance in Table 7.1, for example, the author's estimate of the "best possible" transcription (derived as explained below) was *00 gewonnen heeft of verloren*, which

[3] The precise etymology of this word is a matter of debate. The verb stem *woon* means "dwell", which is not obviously related to the meaning of the word *gewoon*. Some scholars suggest that the word is instead derived from the phonetically-similar verb stem *wen*, which means "become accustomed to".

consists of 10 syllables. Therefore, in this example, all listeners yielded the same intelligibility measure of 0.4. The rationale for setting the syllable count at a fixed value was as follows. Based on the transcriptions, it seemed that some of the listeners had either (a) not invested the time or effort necessary to count syllables with a reasonable degree of accuracy, (b) not fully understood this component of the task, or (c) found syllable-counting based on a limited number of listening occasions to be too challenging. The implications of these deficiencies for future research and for clinical practice are discussed in Section 7.4.3. For the present purposes, the important point is that the listeners' syllable counts clearly had both low validity and low reliability. This was evident from the fact that the number of zeroes used to denote unintelligible word-groups sometimes varied substantially between listeners and/or was markedly different from the value estimated by the author. The syllable count derived by the author was based on a labour-intensive, intricate process that may not be practical for use in the clinic (or even in future research studies with larger sample sizes), but resulted in a value that would have been considerably more accurate than one based on listener judgments, even if the appropriate training had been provided. The procedure involved listening to each utterance, as well as to the monologue as a whole (i.e., without pausing between utterances), on multiple occasions, while also examining all of the listeners' transcriptions. For some utterances, the author also consulted with a native Dutch speaker in order to decipher some of the words that were of borderline intelligibility. This resulted in the production of a "best possible" transcript, as was illustrated for the above example: *00 gewonnen heeft of verloren*. The process was iterative; i.e., an approximate transcript was created to begin with, and it was continually refined on further listening occasions until no additional changes were made. As mentioned, this was a labour-intensive process (which took several hours per monologue), by the end of which the author was highly familiar with all the utterances and could reliably recall their rhythms, including in the unintelligible portions. For this reason, it was not possible to determine a measure of intra-rater reliability for the syllable counts.

Further discussion of the benefits and limitations of the above correction procedures is provided in Sections 7.4.3 and 7.4.4. However, it is worth emphasising at this juncture that the correction procedures were considered to result in substantial improvement to the accuracy of the technique, at least in the present study where the validity of the raw data was in question. This is because the consensus approach allowed for the identification of portions of speech that can be considered intelligible by some objective measure, while the author's syllable count was likely to be far more accurate than would normally be achieved in a perceptual assessment, due to the considerable time and effort that was invested into the judgments, as well as the integration of information from multiple sources.

The previous paragraphs described how the measure of spontaneous-speech intelligibility was derived for each utterance, for a given listener. This procedure was then repeated for every utterance in the monologue to obtain a final measure of intelligibility of the monologue for the listener in question. Ideally, intelligibility scores for two or three different monologues would have been obtained per speaker, to gain a solid understanding of intra-speaker variability. Indeed, multiple monologues were recorded with this purpose in mind for all but two of the speakers (who were too fatigued to continue after the first monologue). However, there were insufficient listeners to enable multiple monologues to be analysed in every case, and hence only four of the speakers in the cohort were assessed using two monologues. These four speakers were chosen either on the basis that their dysarthria was relatively severe or that, to the best of the author's judgment, they produced monologues that seemed to vary markedly in terms of intelligibility. In other words, the aim was to gain a preliminary idea of the *worst-case* consequences of measuring intelligibility based on a single spontaneous-speech sample. For speakers who produced more than one monologue, but where only one monologue was analysed, the *initial* monologue was selected. It was reasoned that this would allow for a fairer comparison (compared to using the second monologue) with the speakers who only produced one monologue, as it is possible that speaker intelligibility changed systematically over time, e.g., due to fatigue and/or the process of becoming accustomed to the task.

7.2.2. Data analysis procedures

The main objective of this study was to examine the correlation between intelligibility in single-word reading and intelligibility in spontaneous speech. To this end, the principal variable of interest for denoting SWR intelligibility was word accuracy (i.e., the percentage of correct words) derived from both orthographic transcription and the multiple-choice study. Word accuracy calculated via these two response modes is a common outcome measure in single-word intelligibility tests,[4] including the Frenchay Dysarthria Assessment (Enderby & Palmer, 2008) and Kent et al.'s (1989) test. Therefore it is important to establish the relevance of whole-word accuracy to intelligibility in spontaneous speech. However, it could also be of interest to investigate other measures of speaker intelligibility derived from single-word reading – in particular, phoneme accuracy and consonant accuracy – as these could show different (and perhaps superior) levels of correlation with intelligibility in spontaneous speech. For example, *phoneme* accuracy might be more indicative of intelligibility than *word* accuracy, as the latter does not differentiate between

[4] Assessment guidelines often stipulate that mild-moderate speakers should be assessed using orthographic transcription while more severe speakers should be judged using an MC protocol.

speakers who tend to yield errors on one phoneme per word and speakers who yield errors on multiple phonemes. Regarding consonant accuracy, McLoughlin (2009) states that it is easy to demonstrate, using sentences in which either all the vowels or all the consonants are replaced with a single, unchanging phoneme, that consonants convey a greater degree of intelligibility than vowels (McLoughlin, 2009: Chapter 3), at least in the English language. Indeed, Flipsen et al. (2005) noted that the percentage of correct consonants (PCC) is a commonly-used metric in intelligibility research, and Lagerberg et al. (2014) used it as their metric of single-word intelligibility for examining the correlation with SSI in children with speech-sound disorder. Shriberg and Kwiatkowski (1982) showed that PCC, calculated from spontaneous-speech samples, is highly correlated with subjective judgments of severity in children with phonological disorder.

The main data analysis procedure, namely to examine the relationship between measures of SWR intelligibility and intelligibility in spontaneous speech, was carried out using Pearson's correlation coefficient (*r*), having first determined that the variables in question were normally distributed (using the Shapiro-Wilk test). In cases where the data were non-normal, Spearman's *rho* was calculated. The evidence in the literature shows that it is highly unlikely that there would be a negative correlation between intelligibility in single-word reading and spontaneous-speech intelligibility. Therefore the level of significance was calculated under the assumption of a right-sided alternative hypothesis (*r* or *rho* > 0).

The second data-analysis procedure examined three characteristics of the monologues that were hypothesised to influence the relationship between SWR intelligibility and intelligibility in spontaneous speech: utterance length, speech rate and fluency. A rigorous analysis of these variables was considered beyond the scope of this study (see Chapter 2, Section 2.3), and consideration of these variables did not inform the study design. Rather, after examining the data depicting the relationship between single-word reading accuracy and SSI (see Section 7.3.1), it was considered worthwhile calculating simple metrics that reflect utterance length, speech rate and fluency, as it was thought that these factors might explain some of the unexpected findings (i.e., speakers who yielded either a higher or a lower SSI than would be expected based on their accuracy in single-word reading). The definitions of the three variables are provided below; firstly, the method of assessing their relationship with spontaneous-speech intelligibility is described.

Ideally, the data would have been analysed by means of a technique such as multiple regression. However, due to the low sample size, it was not justified to employ a method that simultaneous assesses the correlation of SSI with multiple independent variables (four in this case: utterance length, speech rate, fluency and a metric of single-word reading

accuracy). In the case of multiple regression, for example, a calculation of Cohen's (1988) effect size shows that the minimum value of R^2 that is capable of being detected based on ten subjects and four independent variables, assuming a power of at least 0.80 and an alpha of 0.05, is 0.751. Furthermore, even if an effect size of this order of magnitude were to be observed, it would be likely to have low validity. Thus the explanatory variables were investigated by means of a series of univariate correlation calculations. As shown in the Results section, a multiple linear regression using just two of the independent variables (SWR accuracy and fluency) was attempted; however, the change in R^2 relative to a simple regression using SWR accuracy alone was found to be insignificant. The decision as to whether to conduct a one-sided or two-sided correlation test was made separately for each explanatory variable, as described in the following paragraphs.

Mean utterance length. A number of studies have investigated the relationship between utterance length and intelligibility in speakers with dysarthria (e.g., Allison et al., 2019; Yunusova et al., 2005). However, previous studies employed different designs from the present investigation; in particular, they generally assessed SSI for reading tasks rather than for natural speech. The study by Yunusova et al. (2005), while it was based on sentence reading, at least separated the speech into breath groups. This is more relevant to the present study than an examination of the correlation between intelligibility and sentence length in a reading task. Yunusova et al. (2005) found that speakers with a higher number of words per breath group tended to have higher intelligibility. Lagerberg et al. (2014) reported a positive correlation ($r = 0.78$, $p < 0.01$) between average utterance length and SSI in children with speech-sound disorder. Tjaden and Wilding (2011) measured the effect of utterance length in speakers with PD on a *within*-subject basis, where an utterance was defined as a stretch of speech bounded by a silent period or pause of at least 200 ms. They found that, both for reading passages and monologues, most speakers showed a positive correlation between the number of words per utterance and the utterance intelligibility, which they interpreted as evidence that contextual cues facilitate intelligibility in dysarthric speech. Based on these findings, it was hypothesised that a positive association would also be observed in the present study; thus a one-sided correlation was performed.

Lagerberg et al. (2014) were able to calculate the lengths of the utterances directly from the edited speech samples presented to the listeners, since the monologues were divided up according to natural semantic pauses. Although the author had intended to use the same methodology in the present study, in practice, some "natural" utterances (i.e., word-groups that did not contain an obvious pause) had to be split into *two* utterances for the purposes of the listening sessions. This tended to occur in the following situations: (a) highly

intelligible, fluent speakers who sometimes produced very long utterances which, in the author's opinion, might have been difficult for the listeners to recall and transcribe; and (b) speakers of low intelligibility for whom, in the author's judgment, it was difficult to count syllables in unintelligible utterances above a certain length. A second complication arose when deciding how to calculate utterance length: in speakers who used relatively large numbers of specialist words or proper nouns, it needed to be decided whether the measure of utterance length would exclude or include such words. If the underlying cause of a positive association between utterance length and intelligibility is increased semantic context, then specialist words should not be counted (in fact, they may even be a hindrance to intelligibility). On the other hand, if the important factor is the natural utterance length produced by the speaker, irrespective of its content, then specialist words should be included. In order to deal with these two complications, *two* metrics of utterance length were devised. The first metric, referred to herein as the "perceived" utterance length, was calculated from the utterances presented to the listeners, which, as mentioned, were sometimes created by dividing up a natural utterance into two parts. Furthermore, this metric excluded specialist words and proper nouns, as these are unlikely to be of value for deciphering the message. The second metric, referred to as the "produced" utterance length, was calculated from the natural utterances produced by the speaker (i.e., utterances parsed on the basis of semantic pauses), and it included specialist words and phrases. In other words, the first metric assumes that the information *conveyed to the listener* drives the relationship, while in the case of the second metric, speaker intelligibility is assumed to be correlated with the speaker's *natural* utterance length.[5] In both cases, the utterance length for a given monologue was calculated as the mean number of syllables per utterance. A metric based on the number of *words*, which has been used by other authors, would not have been feasible due to the fact that word boundaries could not always be determined.

Speech rate. It is well known that deliberately reducing the speaker's rate of articulation can improve connected-speech intelligibility in people with dysarthria, although there are also individuals who seem to show the opposite trend (van Nuffelen et al., 2009a). In cases where speech rate is determined by the speaker (rather than being externally controlled), both outcomes seem possible. If the population consists of speakers with an approximately equal level of impairment, then the highest intelligibility levels might be exhibited among the speakers who are most successful at reducing their speech rate (i.e., a negative association). Alternatively, if the speakers vary widely in terms of their level of impairment, then it might be the case that the speakers who are more severe are more likely to attempt

[5] For some monologues produced by some speakers, the two metrics did not differ.

to compensate by using slowed speech. Assuming that full compensation cannot be achieved, this would lead to a positive association between speech rate and intelligibility. A positive association could also arise naturally in the uncontrolled scenario, as speakers with a more severe dysarthria might have lower respiratory reserve and an increased propensity for articulatory fatigue, leading to a reduced speech rate. This is especially likely if their dysarthria is caused by a progressive disease such as ALS (Allison et al., 2019). Given these arguments, no prediction was made regarding the direction of the association between speech rate and intelligibility (a two-sided test).

Speech rate was calculated as the number of syllables uttered per minute, which was assumed to be an indication of the speed of movement of the articulators. It was calculated by dividing the total number of syllables in the monologue (as determined by the author; see Section 7.2.1) by the sum of the durations of the individual utterances. In other words, silent pauses between utterances, which were found to vary considerably in duration among speakers, were eliminated. This was reasoned to be a more reliable measure of the speed of movement of the articulators than a metric that included such pauses. Ideally, *within*-utterance pauses would also have been excluded, but the identification of such pauses would have been a laborious task and was considered beyond the scope of the study.

Fluency. Dysfluencies such as hesitations, pauses and repetitions may have an important influence on spontaneous-speech intelligibility in individuals with speech impairment (Miller, 2013). In addition, speakers with more severe dysarthria may be more likely to exhibit pausing behaviour, especially when dysarthria is a consequence of a progressive condition (Rong et al., 2016). Both of these mechanisms imply a positive association between the level of fluency and spontaneous-speech intelligibility. Therefore, this variable was assessed using one-sided correlation.

A simple measure of fluency was derived, namely the percentage of the speaking time occupied by silent pauses between utterances – thus, smaller values indicated greater fluency, such that a negative correlation with SSI was expected. The pauses between utterances were not presented to the listeners and thus could not have had a *direct* effect on SSI. However, it was hypothesised that a speaker's between-utterance pauses depended on factors such as their cognitive state, level of fatigue, prosodic skills, and language abilities,[6] and that these attributes would have also affected their *within*-utterance pausing behaviour, as well as other important features of the monologue, such as syllable stress

[6] Note that mild aphasia may have been present in some speakers.

patterns, communicative vividness and discourse coherence. Ideally, the dysfluency metric would have also reflected within-utterance features. However, in addition to the fact that the analysis of within-utterance dysfluencies is a laborious process, in the present study it was further complicated by the fact that the target stimulus was not known. Therefore, it was often very difficult (if not impossible) to differentiate between sounds that could be considered dysfluencies (e.g., restarts, fillers, word fragments and self-corrections) and sounds that were an attempt to convey unique information, but were unintelligible.[7] Thus an in-depth analysis of the association between intelligibility and pausing behaviour was left for future research.

To obtain the dysfluency metric, it was not the pauses themselves that were measured, but the durations of the utterances. The beginning and end points of each utterance were identified by examining various features of the waveform and its spectrogram in Praat, including the intensity contour, the formant contours and the amplitude of the sound signal. These sources of information, along with listening to the utterance in an attempt to identify the first and last phoneme, were integrated to produce the author's best possible subjective assessment of the start and end points of the utterance. This was sometimes challenging, as shown in Figs. 7.1 and 7.2. Nevertheless, it was estimated that the typical worst-case level of uncertainty on the identification of the transition point, expressed as the difference between the most conservative estimate and the most lenient estimate, was 0.05 s. Even if this error is assumed to apply to both the start point and end point of every utterance in the monologue, and assuming that the monologue consists of 12 utterances, this would have resulted in a maximum uncertainty in the final outcome measure (the percentage of the monologue occupied by between-utterance pauses) of just ± 5%.[8]

Having calculated the durations of the individual utterances in a monologue, these were summed and subtracted from the total monologue duration (i.e., the time period between the start of the first utterance and the end of the last utterance). This resulted in a measure of the total length of time occupied by between-utterance pauses, which was expressed as a percentage of the total monologue duration.

---

[7] This may be an issue of limited resources. In other words, it might be *possible* to make such distinctions with a reasonable degree of accuracy, but it would have required many hours of analysis per monologue, carried out in conjunction with a fluent Dutch speaker who is familiar with the Antwerp accent.

[8] The lowest value for the outcome measure (dysfluency) observed in the cohort was of the order of 6%. Thus a percentage uncertainty of 5% would mean that the true value could have ranged from 5.7% to 6.3%. For the highest observed dysfluency value (~36%), the corresponding range is 34.2% to 37.8%.

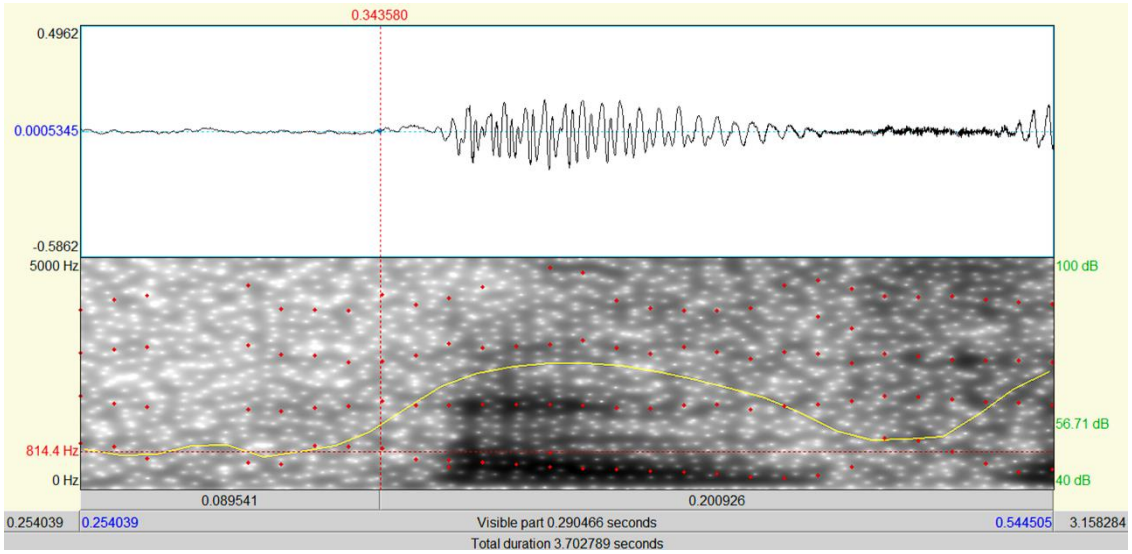


Figure 7.1. Example of a case where it was difficult to locate the precise start-point of the utterance. The initial phoneme could not be identified, but it was thought to be a vowel. The yellow trace represents the intensity contour. The vertical red dotted line indicates the chosen transition point.

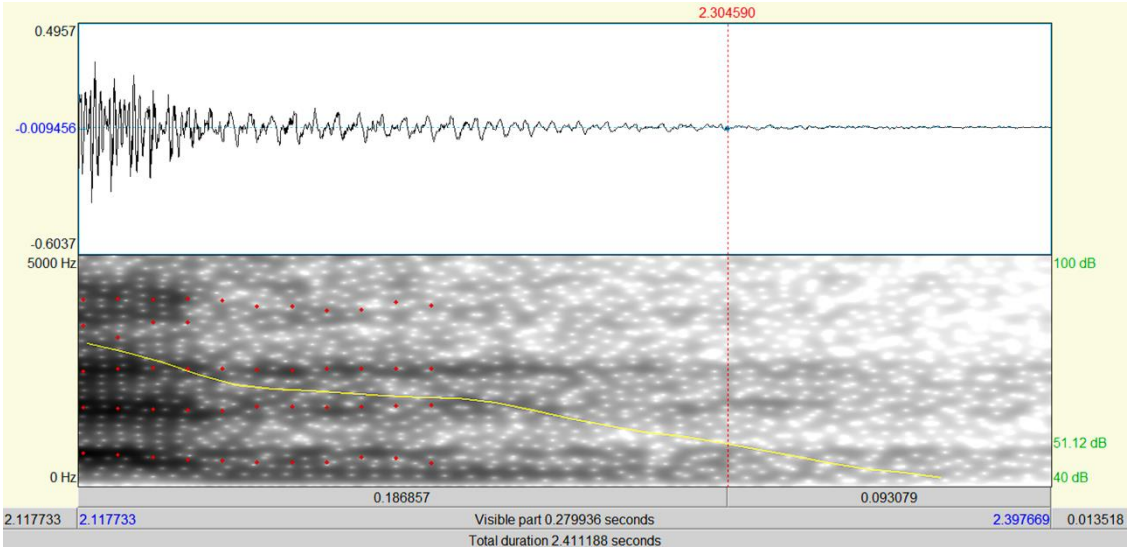


Figure 7.2. A case where it was difficult to locate the end point of the utterance. The transition was from a schwa to silence. The vertical red dotted line indicates the chosen transition point.

7.3. Results

7.3.1. Correlation between intelligibility in single words and in spontaneous speech

Table 7.2 shows the spontaneous-speech intelligibility results for all the monologues analysed in the study. For all but one monologue, the mean intelligibility value (± 1 SD) was calculated by averaging across the set of listeners (between 3 and 5). The remaining monologue, produced by Speaker 6, was deemed to be almost completely intelligible by the

author and indeed yielded an intelligibility score of 97.9% from the first listener.[9] Therefore, the monologue for this speaker was not subjected to further assessment. The second column shows the English translation of the title (context) of the monologue as presented to the listeners. The number in brackets shows the "familiarity rating" (on a scale of 1-4) as judged by the author. This rating is intended to provide an indication of how familiar the average listener might have been with the subject matter. For example, the first monologue delivered by Speaker 5, which was on the subject of architecture and design, contained references to designers who would not be widely known among lay listeners (such as Kurt Naef), as well as to specialist topics such as the Bauhaus movement.[10] Therefore the familiarity of this monologue was rated at the lowest possible value of 1. The final two columns show word-accuracy scores from single-word reading as assessed via free and forced-choice recognition. As explained in Chapter 6, for two of the speakers, no multiple-choice assessment of their single-word stimuli was carried out. For every monologue, the SSI value, which is a measure of syllable accuracy, is higher than the word-accuracy score achieved by the same speaker in orthographic transcription. The range of SSI values is 67.0% to 97.9%, with a mean of 87.4% ± 8.32% (1 SD). Before calculating the mean and standard deviation for the cohort, the spontaneous-speech intelligibilities for Speakers 4, 5, 9 and 10 were defined as the average SSI for their two monologues.

Figures 7.3a and 7.3b show the correlation between word accuracy and spontaneous-speech intelligibility when word-accuracy scores were obtained from the free- and the forced-choice studies, respectively. As was the case for all correlation calculations in the Results section, when a speaker was judged on the basis of *two* monologues (Speakers 4, 5, 9 and 10), spontaneous-speech intelligibility was calculated as the average of the two SSI scores. When comparing the correlation values in Figs. 7.3a and 7.3b, it needs to be remembered that there are missing data for Subjects 1 and 10 in the multiple-choice mode. However, if the latter two subjects are removed from the analysis for the free-response mode, then the correlation coefficient remains approximately the same (r = 0.60), albeit with borderline significance (p = 0.06, one tail). The difference between the correlation scores for the two response modes for these 8 speakers (r_{free} = 0.60; r_{MC} = 0.69) was non-significant (p = 0.81) using a Fisher's r to z transformation.

[9] The numerator could not be calculated using the consensus approach, as there was only one listener. However, since the monologue was highly intelligible, it was reasonable to assume that all the words transcribed by the listener were indeed "intelligible".

[10] Specialist words and proper nouns were excluded from the calculation of the SSI metric; however, if such words appear frequently, it suggests that the monologue may be of low familiarity, which could reduce its intelligibility.

ID	*Title of monologue (familiarity rating)*	*SSI (%): mean ± 1 SD*	*Word accuracy: free (%)*	*Word accuracy: MC (%)*
1	Christmas (4)	86.4 ± 3.1	78.0	N/A
2	Going on holiday (3)	88.1 ± 2.4	77.3	85.5
3	A difficult period (3)	88.1 ± 4.6	73.2	83.0
4	My former job (4)	87.4 ± 5.2	49.1	73.7
4	My social circle (4)	85.5 ± 0.9	49.1	73.7
5	Architecture/design (1)	88.1 ± 1.6	70.6	90.8
5	My first job (2)	94.2 ± 1.2	70.6	90.8
6	My hobby (4)	97.9 †	83.9	93.4
7	My illness (4)	96.8 ± 0.5	88.4	92.7
8	At the hospital (4)	74.1 ± 10.4	68.8	85.9
9	Family (2)	79.0 ± 2.7	62.0	72.7
9	My grandson plays football (3)	67.0 ± 6.2	62.0	72.7
10	My hobby (2)	93.2 ± 1.8	90.7	N/A
10	How I discovered my hobby (3)	91.2 ± 2.0	90.7	N/A

† No standard deviation available, as only one listener assessed this monologue.

Table 7.2. Characteristics and intelligibility scores of all the monologues assessed in the study. The final two columns show the word accuracies obtained in single-word reading (free-response and multiple-choice). Alternate speakers have been shaded differently (grey vs. white).

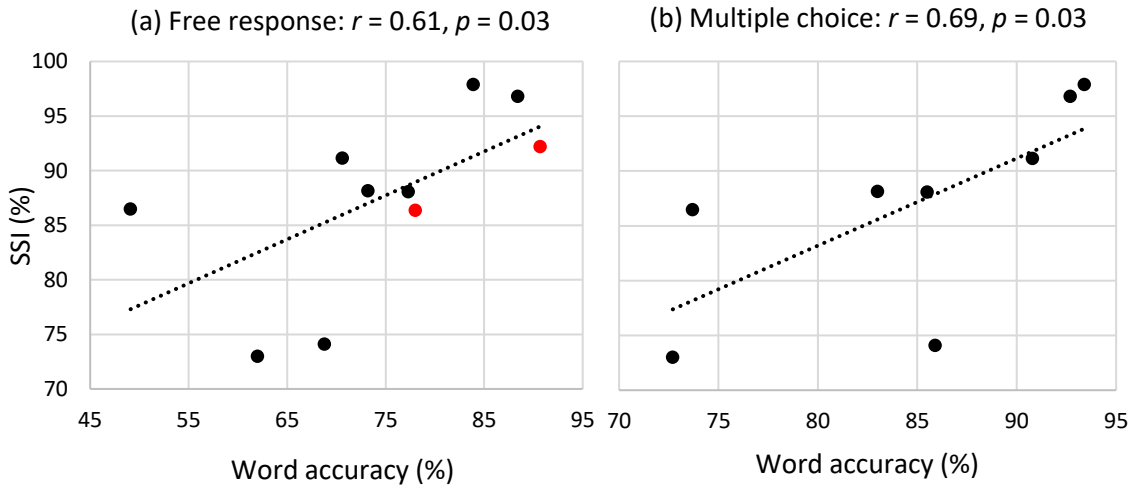


Figure 7.3. Pearson's *r* and one-sided *p*-values for the relationship between SSI and word accuracy calculated from (a) the free-response mode (n = 10) and (b) the MC mode (n = 8). The red circles in the left-hand figure represent speakers who were not assessed in the forced-choice mode.

Figure 7.4 examines the relationship between spontaneous-speech intelligibility and four different accuracy metrics derived from orthographic transcription of the single-word reading stimuli: word accuracy (i.e., a repeat of Fig. 7.3a, to facilitate comparison with the other metrics), phoneme accuracy, C1 accuracy and vowel accuracy. The data for the correlation of spontaneous-speech intelligibility with C2 accuracy are not shown, as Spearman's rank coefficient failed to reach significance (*rho* = 0.52, one-tailed $p = 0.063$). Taken as a whole, the results show that, contrary to expectations, the correlation with *phoneme* accuracy was weaker than the correlation with *whole-word* accuracy. Analysis in terms of the individual segments (C1, V and C2) demonstrates that the lower-than-expected correlation with phoneme accuracy was probably due to the fact that C2 accuracy was poorly correlated with spontaneous-speech intelligibility. While a Spearman's *rho* of 0.52 may not seem that low, this correlation metric does not assume a linear relationship. In fact, the relationship between C2 accuracy and SSI was observed to be far from linear (see Section 7.4.1 for an explanation). Therefore, since C2 accuracy contributes to overall phoneme accuracy, it has the effect of reducing the correlation between phoneme accuracy and SSI. The strongest correlation was found for C1 accuracy ($r = 0.76$, one-tailed $p < 0.01$), providing further justification for the common practice of using PCC as a measure of intelligibility. To facilitate discussion of these findings (see Section 7.4.1), Fig. 7.4 identifies three speakers who (a) can be regarded as outliers in the sense that their SSI value is either higher or lower than would be expected based on their intelligibility in single words (i.e., they are far from the trendline) and (b) were likely to have exerted a strong influence on the calculated correlation coefficient, due to the paucity of data in that region of the graph.

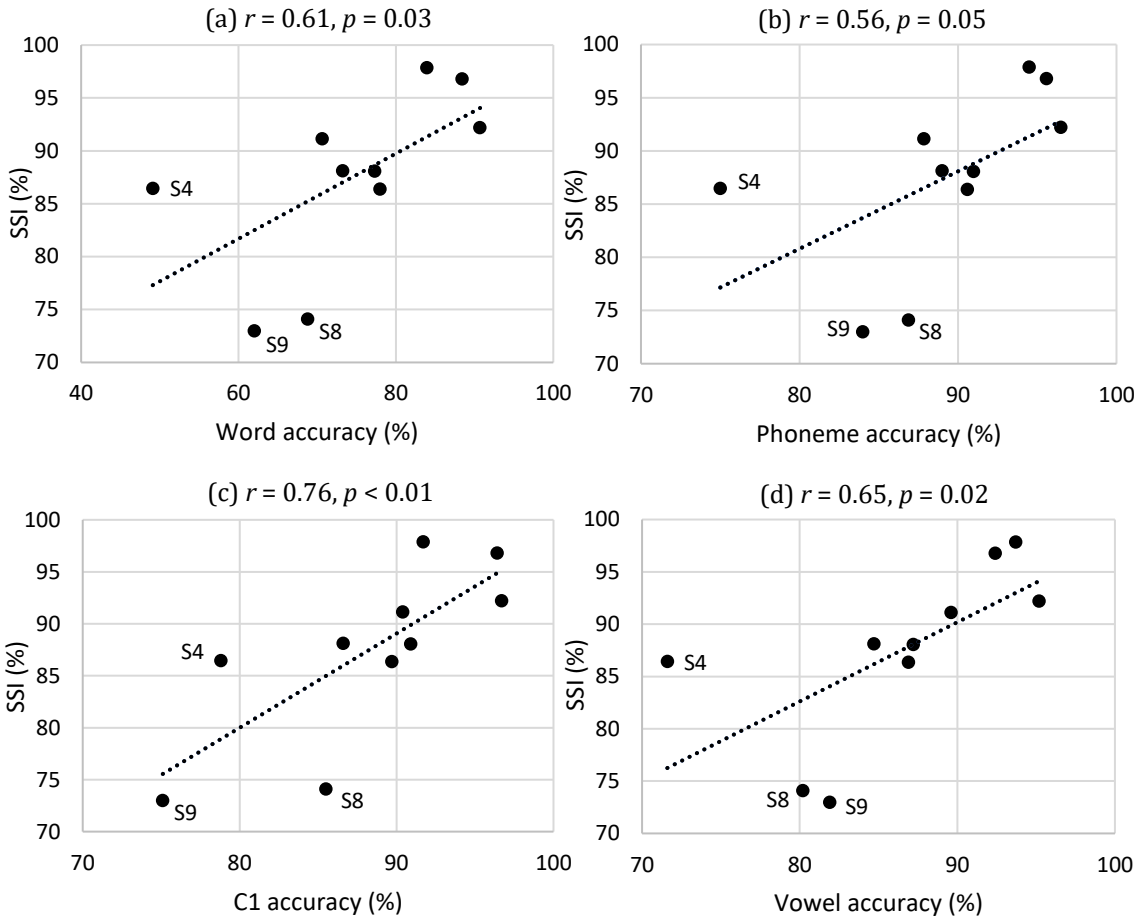


Figure 7.4. Correlation between SSI and four different accuracy metrics from the orthographic transcription of single words (see x-axis labels). The datapoints of three speakers (S4, S8 and S9) have been labelled, to facilitate the discussion in Section 7.4.1. The dotted line shows the best-fit linear regression, the slope of which is equal to Pearson's *r*.

7.3.2. Other explanatory variables

Table 7.3 presents the findings for the correlations between SSI and the three explanatory variables. For the speakers judged on the basis of two monologues, a single datapoint was created by averaging over the two SSI values as well as over the two utterance-length (or speech-rate or dysfluency) values. This ensured that the 10 datapoints used in the correlation analysis would be statistically independent. As explained in Section 7.2.2, utterance length was calculated in two ways: the number of syllables per "produced" utterance and the number of syllables per "perceived" utterance. The latter showed a stronger association with SSI and was therefore used in subsequent analyses. To check for collinearity, the variance inflation factors (VIFs) were calculated for a correlation matrix of the three variables (speech rate, dysfluency and perceived utterance length). This revealed no concerns about collinearity (maximum VIF of 1.36), meaning that the three variables can be considered to be reasonably independent.

Variable correlated with SSI	*Pearson's r*	*p (one or two tailed)*
Produced utterance length (syllables per utterance)	0.30	0.20 (one tailed)
Perceived utterance length (syllables per utterance)	0.44	0.10 (one tailed)
Speech rate (syllables per minute)	-0.33	0.35 (two tailed)
Dysfluency (% duration of monologue occupied by between-utterance pauses)	-0.58	0.04 (one tailed)

Table 7.3. Results of correlation analysis between spontaneous-speech intelligibility and three different characteristics of the speakers' monologues: utterance length, speech rate and dysfluency.

Table 7.3 shows that, as predicted, perceived utterance length was positively associated with SSI, while dysfluency exhibited a negative correlation. However, only the dysfluency finding is significant (for an α level of 0.05) and it becomes insignificant if a Bonferroni correction is applied ($\alpha = 0.05/3 = 0.017$).[11] Speech rate showed a weak negative correlation, but this was far from significant. Inspection of the speech-rate data (not shown) revealed that a prominent outlier was Participant 9, who despite having the fourth slowest speech rate was the least intelligible speaker in spontaneous speech. This speaker, who had experienced a cerebellar stroke, seemed to exhibit "excess and equal stress", a supposed hallmark of ataxic dysarthria (Duffy, 2005: Chapter 6). As discussed further below (see Section 7.4.1), Participant 9's spontaneous speech also stood out for other reasons. He was the only speaker who seemed to produce segmental distortions and substitutions of a relatively consistent nature. Moreover, his propensity to make segmental errors, as well as the other perceptual characteristics of his speech, were highly consistent from one utterance to another. In short, his intelligibility did not seem to fluctuate within or between utterances due to factors such as speech rate or level of effort; rather, it seemed to be fundamentally limited by his production impairments. Nevertheless, even if the datapoint for S9 is omitted from the analysis, Pearson's *r* only increases to -0.50 (two-tailed $p = 0.17$). Therefore, speech rate does not seem to be strongly associated with SSI in this study.

The correlation with dysfluency was marginally significant and therefore warrants special attention. Figure 7.5 shows the dysfluency data in graphical form. It can be seen that, once again, the most outlying datapoint is that of Speaker 9, whose SSI value is lower than would be expected based on the trendline. A multiple linear regression was conducted using

[11] This correction is likely to be overly conservative, as the three tests are not fully independent.

dysfluency and C1 accuracy as predictor variables, where the latter metric was chosen because it produced the highest correlation with SSI of all the accuracy metrics calculated for single-word reading. The outcome was compared with that of a simple linear regression using C1 accuracy only. There was a small increase in adjusted R^2 for the more complex model (0.567 relative to 0.523). However, an R^2-change F-test showed the level of fit improvement to be insignificant (p = 0.22). Furthermore, the standardized beta coefficient for dysfluency (-0.324) in the multiple-regression model was non-significant (t = -1.35, p = 0.22).

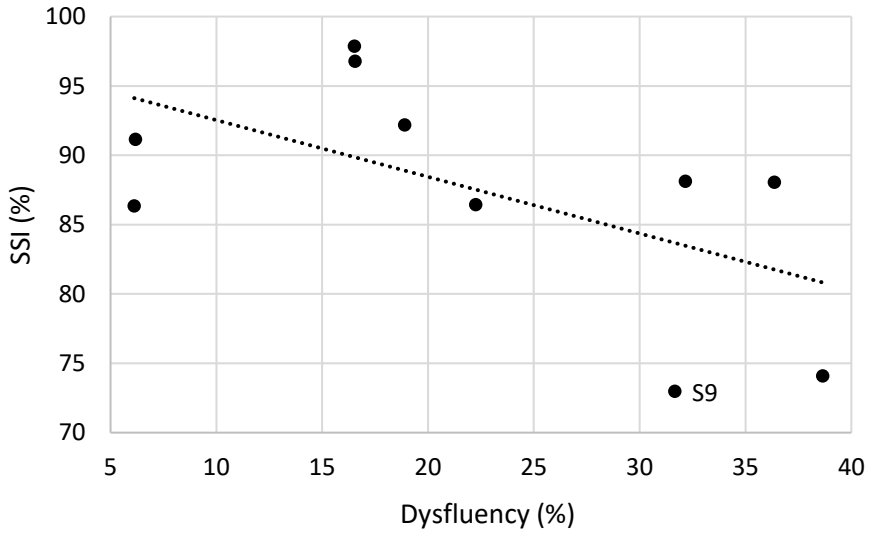


Figure 7.5. Correlation between dysfluency (the proportion of the duration of the monologue occupied by between-utterance pauses) and SSI (r = -0.58, one-tailed p = 0.04).

7.4. Discussion

7.4.1. Relationship between intelligibility in single words and in spontaneous speech

For every speaker, irrespective of their level of severity, the measure of syllable accuracy derived from spontaneous speech was higher than the corresponding word-accuracy value obtained in the single-word reading task (judged via orthographic transcription). This is consistent with the findings of Hustad (2007) for speakers with cerebral palsy. Specifically, Hustad only observed an improvement in all speakers when the sentences formed a *narrative*, whereas the difference in intelligibility between single words and unrelated sentences was small or insignificant in speakers with moderate and severe dysarthria. Improved intelligibility in a narrative is due to the availability of additional context and cues. It appears that these cues more than compensate for any loss of intelligibility due to phonetic reductions in unstressed portions of speech or the extra burden on the speaker of having to produce connected speech. In addition, in the present study, the single words

were chosen to be minimally contrastive with a large number of real words, thereby encouraging errors in this task.

The highest correlation between single-word reading intelligibility and SSI, which occurred when SWR intelligibility was defined as C1 accuracy, was $r = 0.76$ ($p < 0.01$). This value is in broad agreement with previous studies, which have reported correlations of $\gtrsim 0.8$ between measures of SWR and connected-speech intelligibility (e.g., Lagerberg et al., 2014; Yorkston & Beukelman, 1978; Yunusova et al., 2005). There would be little purpose in directly comparing correlation values across these studies, as there are substantial differences in study design. Furthermore, due to the low sample size in the present study, the correlation values were highly dependent on the specific data – in other words, removing or adding just one datapoint could result in a substantial change to Pearson's *r*. The difference between the correlation value for C1 accuracy ($r = 0.76$) and that for vowel accuracy ($r = 0.65$), assessed using the Fisher *r*-to-*z* transformation, was not statistically significant ($p = 0.68$, two-tailed test). It was argued above (see Section 7.2.2) that consonants convey a greater degree of intelligibility than vowels in connected speech, at least for the English language. By conducting a similar thought experiment in Dutch,[12] it seems reasonable to make the same assertion. Therefore, although the superior correlation for C1 accuracy (relative to vowel accuracy) did not meet statistical significance in this study, one would expect future studies with larger sample sizes to replicate this result. The reason for the poor correlation with C2 accuracy is that, with the exception of the two speakers who exhibited 'final singleton → cluster' errors on a relatively consistent basis (S4 and S5), the word-final consonants were relatively robust. Thus the range of C2 accuracy values in the remaining 8 speakers was narrow (94.0 – 98.4), meaning that a strong correlation with SSI would be highly unlikely. Finally, the correlation coefficient was marginally higher for word accuracy in the multiple-choice mode than for word accuracy in orthographic transcription ($r = 0.69$ vs. 0.60; n = 8). This difference was not significant; however, a higher correlation for the MC mode would be logical on the grounds that the majority of misarticulations in dysarthria seem to be distortions rather than substitutions (see Chapter 2). It was argued in Chapter 6 that when a phoneme is distorted towards another phoneme, but does not cross the phoneme boundary, the MC mode creates a level playing field in the sense that the listener is equally likely to consider the target as the distractor. In contrast, in a free-response mode, a given distortion may be perceived as a substitution in cases where the minimal-pair distractor has a greater *a priori* chance of

[12] As mentioned in Section 7.2.2, the experiment involves replacing all the vowels in a sentence with one specific vowel phoneme and then repeating the procedure for consonants. It is immediately apparent that there is a greater reduction in intelligibility when the consonants are removed.

being considered, e.g., because it has higher lexical frequency than the target. Spontaneous speech may have a similar effect as the forced-choice mode in terms of reducing the likelihood that subtle distortions will be perceived as substitutions, because the additional linguistic and phonetic cues increase the prior probability of hearing the target.

There are two main mechanisms that could result in low or moderate correlations between single-word reading intelligibility and SSI when measured using Pearson's *r*. Firstly, there are likely to be many factors that contribute to spontaneous-speech intelligibility other than articulatory precision. Secondly, Pearson's *r* assumes a linear relationship between the two variables, which may not reflect reality. From inspection of the graphs in Fig. 7.4, there is no clear evidence of a nonlinear relationship, although there are insufficient data to draw any definitive conclusions. Furthermore, the degree of nonlinearity is likely to depend on the range of severity levels (see Chapter 2, Section 2.3), such that a nonlinear relationship is more likely to emerge in studies where there are speakers at both ends of the spectrum.

Three speakers were singled out in Fig. 7.4 as "outliers" who were likely to have exerted a strong influence on the calculated correlation coefficient (i.e., the slope of the best-fit line). Therefore, it is worth considering whether there were any obvious differences between these speakers and the rest of the cohort, not only as means of understanding why they are outliers, but also to establish whether their contribution to the trendline can be considered justified. Speakers 8 and 9 had considerably lower SSI values than the remainder of the sample and, in general, their spontaneous-speech intelligibility was lower than would be expected based on the trendline. Consider, first of all, Speaker 8. He was a gentleman with a right-sided, ischemic, cortical watershed lesion. He was identified as having mild cognitive impairment, and during the monologue (where he talked about his hospitalisation), he seemed to show emotional lability that translated into fluctuating intelligibility. More specifically, when he became emotional, his speech became very fast, high-pitched and reduced, resulting in a considerable loss of intelligibility. He did not seem to show much awareness of his listener; for example, he did not check for signs that his message had been understood, nor did he appear to invoke any strategies to compensate for his impaired speech. Therefore, the fact that Speaker 8 yielded a lower SSI value than expected (given his single-word reading accuracy) is not surprising, and is likely to be a consequence, at least in part, of his co-occurring cognitive-communication difficulties. If the data were to be re-analysed without this speaker, which could be considered justified given the aforementioned deficits, then the correlation coefficient for C1 accuracy would increase from 0.76 (p = 0.006) to 0.83 (p = 0.003).

Speaker 9 experienced a cerebellar CVA almost 7 years prior to the interview that left him with classical cerebellar symptoms such as an ataxic gait and "scanning" speech (excess and equal stress). He was the only participant in the study whose speech could be obviously described in this manner (based on an informal perceptual assessment carried out by the author; see Appendix 4). Therefore, Speaker 9's prosodic deficit may have contributed to his reduced SSI. In addition, he used frequent proper nouns in his monologues (names of people and places), which were often transcribed with zeroes (i.e., "unintelligible") by the listeners. For this reason, the familiarity ratings of Speaker 9's monologues (see Table 7.2) were lower than would be expected given the mainstream subject matter. In the author's estimation, however, the main barrier to the intelligibility of this speaker was the fact that substitution errors of a fairly consistent nature could be clearly identified in his spontaneous speech, a finding that did not seem to apply to the rest of the cohort (see Appendix 4). These substitutions generally matched the speaker's perceived errors in single-word reading, such as /l/ → /r/ substitutions within consonant clusters (e.g., /blɛi/ → [brɛi], /klɛin/ → [krɛin]) and the distortion of /e/ so that it more closely resembled /ɑ/. The observation that this participant makes a higher number of substitution (as opposed to distortion) errors than his peers is consistent with earlier findings in the thesis; he yielded the highest level of inter-listener agreement in the MC study (Table 6.3) and the second highest level in the orthographic-transcription study (Table 4.7). Furthermore, Table 6.2 reveals that he showed considerably less improvement in word accuracy between the free- and forced-response modes than did other speakers of a similar severity level. As discussed in Chapter 6, all of these findings are likely to be indicative of a high ratio of substitution to distortion errors, at least relative to other speakers in the cohort. Despite these arguments, it is important to point out that in Fig. 7.4c, where the single-word intelligibility metric is *C1 accuracy*, the datapoint of Speaker 9 is no longer an outlier.

The remaining participant worthy of discussion is Speaker 4. This lady had previously undergone surgery for a hemangioblastoma in the fourth ventricle of the posterior fossa. T2-weighted MRI showed residual hyperintensity in the posterior cerebellum and she had symptoms indicative of damage to some of the cranial nerves, including the left-sided hypoglossal nerve. She seemed to have excellent cognition and was a lively and eager communicator. Her delivery was fluent and she had no obvious prosodic deficits. Both of her monologues were given the highest possible familiarity rating of 4. All of these factors are likely to have contributed to the fact that her SSI value was higher than might be expected based on her intelligibility in single-word reading. Nevertheless, the difference in intelligibility between single-word reading and spontaneous speech for this participant seemed remarkable, so to investigate the matter further, a perceptual assessment of her

monologues was conducted by an experienced phonetician who is an expert in the Antwerp accent. He remarked that (Jo Verhoeven, personal communication) "the breakdowns are very localised, i.e., it is not the case that every word or phrase has a breakdown, so there is a large amount of contextual information. In the first [monologue], there are almost no substitutions. There are a couple of deletions which are very common in regional Dutch and that native speakers are familiar with. The real pronunciation problems are imprecise articulations, which are not so difficult to deal with perceptually." According to this account, it would seem that the large number of phonemic substitutions perceived in single-word reading for this speaker were not observed in her spontaneous speech.[13] Several explanations can be imagined, using arguments based on both production and perception. For example, it could be the case that Speaker 4's articulatory precision actually *improves* in connected speech, perhaps indicating that she has initiation difficulties. It is also possible that certain types of error are more likely to be produced in single-word reading, such as /r/ → fricative substitutions, as a consequence of the stressed environment or of the speaker's propensity to hyper-articulate.[14] Therefore, the incidence of such errors might be expected to decline in spontaneous speech where there are unstressed syllables and where the tendency to hyper-articulate is reduced (Lindblom, 1990). Another possible explanation is that Participant 4's misarticulations primarily consisted of distortions. As shown in Chapter 6, there are several mechanisms by which distortion errors might be coded as substitutions in the orthographic transcription of single words. It is hypothesised that the perception of distortions as substitutions would be considerably less likely in spontaneous speech. The fact that Speaker 4 showed a substantial increase in word accuracy (25%) between the free- and forced-response modes could be regarded as evidence in support of the suggestion that she produced a large number of distortion errors relative to other participants.

More generally, as mentioned above, an informal perceptual assessment revealed that substitution errors were not often observed in the spontaneous speech of any of the participants, with the exception of S9. This perhaps calls into question the assumption that articulatory treatment strategies based on substitution errors observed in single-word reading will produce a worthwhile improvement in intelligibility in spontaneous speech.

[13] An exception to this statement was 'final singleton → cluster', which was perceived consistently (see Appendix 4).

[14] It was suggested in Chapter 4 that speakers might use frication as a means of reinforcing the level of articulation when unable to produce a trill. In spontaneous speech, if these attempts are abandoned and the alveolar trill is simplified, e.g., to an alveolar tap, this could be interpreted as a legitimate allophone of /r/.

Nevertheless, the fact that segmental substitutions are not heard in the *intelligible* portions of speech does not necessarily rule out the possibility that they are the main cause of *breakdowns* in intelligibility. Further research would be required to examine this question. In particular, such research would ideally involve obtaining a transcript of the speaker's monologue (which could be provided by the speaker themselves) in order to study the differences between the intended and the perceived utterance. In the absence of information about the intended target, a detailed analysis of the intelligibility breakdowns in the present study is clearly limited. However, the most common observation made by the author was that breakdowns in intelligibility occurred within portions of speech that were fast, unstressed and produced with comparatively low effort. Often, these were parenthetical elements that were not crucial to the narrative flow. In these segments, the speech signal was characterised by reductions, weakened articulations, deletions and telescoping. The general impression was that the speech sounded "slurred". Such a description is rather non-specific and does not amount to much more than saying that the speaker has dysarthria. It is akin to assigning high ratings to Darley's perceptual dimensions of "imprecise consonants" and "distorted vowels" – a practice that was critiqued at the beginning of this thesis, as it does not lend itself to the development of a tailored treatment programme. Thus, there is a great need for further research that aims to improve our understanding of the relationship between segmental speech errors and spontaneous-speech intelligibility. Some potential directions are suggested in Chapter 8.

7.4.2. Correlation between SSI and the explanatory variables

The purpose of calculating the correlation between SSI and the three explanatory variables (utterance length, speech rate and dysfluency) was to determine whether any of these parameters might be able to shed light on the outlying data in Fig. 7.4. However, all three variables exhibited only a weak correlation with SSI, and when an attempt was made to include them in a multiple-regression model, there was no improvement in the fit relative to a simple regression based on C1 accuracy alone. Despite the fact that these parameters were not useful in understanding the variability in spontaneous-speech intelligibility, it is worth briefly discussing whether any insights can be drawn from the present findings.

Utterance length. Lagerberg et al. (2014) identified two different (but not mutually exclusive) mechanisms for the positive association between utterance length and intelligibility: (1) A greater utterance length brings more context to individual words, in line with top-down theories of speech perception; (2) Speakers who produce longer utterances are more talkative. Such individuals are also likely to be more engaged and vivid in their expression and thus might use additional (e.g., prosodic and linguistic) cues. For

adults with dysarthria, a third mechanism can be imagined, namely: (3) Speakers who have a lower level of impairment are less affected by articulatory fatigue and low respiratory reserve. Therefore, they are able to produce longer utterances. Thus, the first mechanism is perceptual, while the second two are related to speech production. In the present study, the positive correlation between utterance length and intelligibility was higher for perceived utterances (r = 0.44) than produced utterances (r = 0.30). This difference was not statistically significant (p = 0.75). However, if it could be observed in future studies, then it would suggest that of the three mechanisms suggested above, the most important is the greater degree of context afforded to the listener by longer utterances.

It seems reasonable to contend that the lower correlation between utterance length and SSI in the present study, compared to the value of 0.78 reported by Lagerberg et al. (2014), is largely due to differences between children and adults with regard to variability in *linguistic* skills. Children with a mean age of 6;0 can vary considerably in their narrative skills, and it is not surprising that utterance length would be correlated with many of the factors that affect discourse comprehensibility in this population. Furthermore, children with speech delay, even if this is their only diagnosis, may also have a degree of language impairment (Binger et al., 2016), which could further increase the variability in narrative skills between subjects. It could *also* be the case in the present study that utterance length was correlated with linguistic factors that affected discourse comprehensibility. For example, it has been shown that some older adults produce shorter utterances with lower syntactic complexity than younger adults, due to a decline in working memory (Kemper & Sumner, 2001). One can imagine that a decline in working memory might affect narrative skills. Furthermore, it is likely that some of the participants had mild aphasia, which, if non-fluent, might have reduced their mean utterance length. Nevertheless, if these mechanisms did play a role, then it was probably minor compared to the role played by linguistic factors in children with speech delay. A further point worth mentioning is that the utterance-length metric employed by Lagerberg et al. (2014) was based on the number of words, rather than the number of syllables. It is possible that a metric based on the number of words would have yielded a higher correlation with SSI in the present study.

Speech rate. The correlation between speech rate and spontaneous-speech intelligibility was weak and nonsignificant (r = -0.33, two-tailed p = 0.35). This could be a reflection of the fact that, as discussed in Section 7.2.2, an association in either direction may be expected, especially in the situation where speech rate is not externally controlled. Thus these competing mechanisms could have cancelled each other out, resulting in no

significant correlation. On an individual level, from listening to the monologues, it was clear that a speaker's level of intelligibility often declined during their faster portions of speech.

Dysfluency. Of the three explanatory variables tested in this study, the crude measure of dysfluency (the proportion of the duration of the utterance occupied by between-utterance pauses) showed the strongest correlation with SSI (r = -0.58, one-tailed p = 0.04). In Section 7.2.2, it was hypothesised that the pauses between utterances, although not heard by the listeners, co-occurred with other factors that *did* affect intelligibility (but would be more difficult to measure). There is some evidence to suggest that therapy aimed at reducing dysfluencies can lead to improved intelligibility in some clinical populations (Miller, 2013). However, even if this is not the case, an improved understanding of the effect of dysfluency could still have clinical benefit, e.g., by allowing prediction of the degree of improvement that might be expected due to articulatory treatment alone.

7.4.3. Suitability of the current technique for quantifying SSI in speakers with dysarthria

The secondary aim of this study was to assess the suitability of applying the method developed by Lagerberg et al. (2014) to adult speakers with dysarthria. Due to the fact that two major modifications were made to the original method, it is not possible to conduct a full and fair assessment. It was hypothesised that the unacceptable amount of guesswork observed in the present study (with respect to both transcribing words and counting the number of syllables in unintelligible portions) was largely a consequence of inadequate training of the listeners and/or a lack of listener effort (especially in the case of online sessions performed by listeners who had no connection with the project). Indeed, it was observed that in the case of syllable counts, the estimates produced in the live sessions, where listeners received a demonstration of the method and the author was present throughout, were more accurate than syllable counts produced online. In addition, some of the online sessions were performed very quickly, suggesting that the listener may not have expended sufficient effort. These observations suggest that the deficiencies of online data collection were at least partly responsible for inaccurate data. Nevertheless, from the author's own experience with syllable counting, it can certainly be concluded that the task is not straightforward, even when one has unlimited time as well as the freedom to repeat, pause and parse the utterances as desired. Connected speech in general, and particularly in speakers with dysarthria, is subject to reductions and deletions, meaning that when speech is unintelligible, the number of perceived syllables could be an underestimation. In addition, in the present study, it was found that in the unintelligible portions of speech, it was not always possible to distinguish between attempts at meaningful words and utterances that can be considered "non-lexical", such as fillers and part-word repetitions

(which are not supposed to be counted according to Lagerberg et al.'s (2014) scoring technique).

In summary, there is insufficient information at present to draw a definitive conclusion about the applicability of the technique to adult speakers with dysarthria. However, it seems likely, first of all, that a minimum requirement for achieving high levels of intra- and inter-rater reliability will be to administer a rigorous training session that includes practice transcriptions, discussion and feedback. If this measure is not successful at achieving acceptable levels of reliability, then it may be worthwhile modifying the technique along the lines of the solutions implemented in the present study – i.e., a consensus approach for identifying intelligible portions of speech and a fixed method of obtaining an accurate syllable count. However, these solutions were extremely time-consuming, so they may need to be automated (or semi-automated) to be practically implementable in future studies, especially those with larger population samples. Second of all, it could prove to be the case that high intra- and inter-rater agreement can only be achieved for a listening population consisting of individuals with at least some formal experience of listening to and assessing disordered speech. In Lagerberg et al. (2014), the listeners were students and graduates of an SLT study programme, and it is possible that this contributed to the greater success of the authors in implementing the technique (relative to the present study where a correction procedure was needed). Previous research on the importance of listener experience when transcribing dysarthric speech has produced mixed findings. Dagenais et al. (1998, 1999) reported that *experienced* SLTs yielded higher intelligibility scores than naïve listeners in the task of orthographic transcription of words and sentences from the Assessment of Intelligibility of Dysarthric Speech (Yorkston & Beukelman, 1981). Smith et al. (2019) found no significant difference in intelligibility scores from trained and untrained assessors for the same task. However, the trained assessors were SLT students with no clinical experience of the population in question (speakers with Parkinson's disease), and the speakers had mild dysarthria, producing a mean intelligibility score in sentence reading of the order of 95% (SD $\simeq$ 6%). Tjaden and Wilding (2011) asked non-expert listeners to carry out orthographic transcription of a reading passage that had been uttered by 12 speakers with PD. Intelligibility scores for the best and worst listener assigned to each speaker differed by an average of 15% (SD = 12%), a finding that the authors described as consistent with other studies showing that non-expert listeners vary in their ability to orthographically transcribe dysarthric speech.

Another factor that needs to be considered when assessing the utility of the technique is the variability in intelligibility scores among different monologues produced by the same

speaker. Due to the low number of listeners, it was only possible to measure inter-monologue variability for four speakers. The average percentage difference between the two intelligibility scores for these four speakers was 5.5%, but the range was 2.0% to 12.0%. Therefore it appears that although, for most speakers, the assessment of intelligibility based on just one monologue would be reasonably reliable, there are individuals for whom this does not hold. Future researchers and clinicians who wish to implement Lagerberg's method should aim to assess intelligibility based on more than one monologue. Variability between monologues is to be expected, not only due to variations in speech-production characteristics resulting from the speaker's level of effort or fatigue, but also due to variations in features of the monologue itself (e.g., familiarity of the subject matter, average lexical frequency, average level of phonetic complexity). The speakers who were assessed using two monologues in the present study were chosen either on the basis that their dysarthria was relatively severe or that, to the best of the author's judgment, they produced monologues that seemed to vary markedly in terms of intelligibility. Interestingly, Speaker 9, who yielded the largest difference in intelligibility between his two monologues (12%), came into the first category. In other words, based on the author's initial impressions, his two monologues did not seem to vary substantially in terms of degree of articulatory effort or level of monologue "difficulty". In fact, they were both on the same subject matter (the speaker's family, in particular his grandson's hobby of football) and were effectively part of the same continuous discourse. In view of the 12% difference in intelligibility for this speaker, his monologues were reviewed again in an attempt to determine the likely cause. As already mentioned, this speaker was relatively consistent in his delivery: he used excess and equal stress and did not seem to vary his level of effort or speech rate. At no point in the interview did he appear to be fatigued. In addition, as stated above, he seemed to be relatively consistent in terms of the nature and frequency of his segmental substitutions and distortions. Therefore, it seems most likely that the difference in intelligibility was due to linguistic factors. The monologue of lower intelligibility, to the best of the author's judgment, seemed to be less coherent (in terms of the flow of the narrative) and contained several utterances of a rather general nature (e.g., "[my family] call me every day") that did not shed any light on the subject matter. In fact, from the listener's responses, it was clear that some of them had not grasped the subject matter, at least not until they were nearing the end of the transcription.[15] If this analysis is

[15] This implies that these listeners either ignored or did not digest the title of the monologue, despite the fact that the demonstration and practice exercises provided monologue titles and explained how to use them. These instructions were also repeated at the start of the "real" transcription.

correct, it reinforces the importance of the role of the *wider* context (i.e., beyond sentence level) in understanding connected speech (Hustad, 2007).

7.4.4. Limitations of the present study

The previous subsection discussed limitations of the technique in general, which may affect its utility as a clinical or research tool. The main limitation of the present study was the inadequate training of the online listeners, which was hypothesised to be the cause of the unacceptable amount of guesswork. As discussed above, this meant that the second aim of the study – to determine the applicability of Lagerberg et al.'s (2014) *original* method to adult speakers with dysarthria – could not be fully addressed; thus it remains a matter for future research. However, aside from not meeting this goal, the lack of training did not, in the author's view, have any major negative consequences for the study, and does not cast doubt on the validity of the findings. This is because a two-fold correction procedure was invoked to deal with guesswork, which resulted in transcriptions of high credibility: (1) A consensus approach was devised to identify the intelligible portions of speech, thereby obviating the need to rely upon the subjective assessment of listeners in deciding whether they were "reasonably certain" of what they heard; (2) An intricate process that integrated information from multiple sources was used to count syllables in the unintelligible portions, resulting in a final estimate that was considered to be of high accuracy. However, there was one drawback of this correction procedure, as demonstrated in Table 7.1. The two listeners who heard the target word *gewonnen* were unlikely to have been guessing after all (see the author's "best possible" transcription in the table caption). However, since the perceived information did not reach the level of consensus, it was disregarded and the intelligibility score for these two listeners was effectively downgraded from 80% to 40%. As a consequence, the mean intelligibility levels in this study were lower than their "true" values. Furthermore, the amount of downgrading may not have been constant as a function of speaker severity, in which case it would have reduced the correlation between intelligibility in single-word reading and in spontaneous speech. Nevertheless, clear instances of downgrading, such as that shown in Table 7.1, did not appear to be common, and no alternative to a consensus-based correction could be envisaged that would still be objective and rule-based. Therefore, the downgrading error had to be accepted as a limitation of the methodology.

Listeners did not always seem to heed the instruction to take account of the monologue heading. The provision of a title was an additional element introduced by the author; i.e., it was not employed in Lagerberg et al.'s (2014) study. It was reasoned that in everyday communicative situations, especially those involving adults, individuals do not tend to

deliver monologues "in a vacuum"; rather, the narrative is normally prompted by a question (e.g., "What are your hobbies?") or some other trigger to broach a particular subject (e.g., the general topic of conversation). In particular, in the case of speakers with severe dysarthria, the value and ecological validity of transcribing a monologue of low intelligibility, in the absence of any context, would be questionable (whereas for speakers of high intelligibility, a title is unnecessary, as the subject matter usually becomes apparent very quickly). Since there was evidence to suggest that some of the listeners did not make use of the monologue title, there would have been variation in the extent to which listeners were aware of the context of the monologue in the case of the more severe speakers. This would have reduced the correlation between SWR intelligibility and spontaneous-speech intelligibility.

Even if it could be assumed that with sufficient training, the listeners would have made full use of the monologue titles, it is still worth considering how this strategy compares with the alternative of fixing the discourse topic. The advantage of the latter strategy is that it would guard against the situation whereby a speaker produces a monologue on a relatively obscure topic, with low-frequency or specialist vocabulary, such that the provision of a monologue title alone is insufficient to provide a "level playing field" with respect to other speakers. It was explained in Chapter 2 that fixing the topic of the monologue was considered imprudent in the present study, owing to the fact that the speakers varied considerably in terms of factors such as their level of cognition and the state of their physical and mental health. In particular, there was concern that fixing the topic would disadvantage speakers with executive dysfunction, cognitive-communication difficulties or cognitive decline, because such individuals might find it difficult to produce a monologue on a prescribed topic that reflects their true abilities (i.e., in terms of factors such as structure, fluency, coherence, vividness and engagement). Furthermore, due to the variability in the speakers' circumstances, there was no obvious choice of subject matter that would be relevant and motivating for all speakers. For example, Speaker 3, who was a hospital inpatient due to new symptoms indicative of the recurrence of a brain tumour, had executive functioning and other cognitive difficulties. Throughout the interview, he demonstrated topic perseveration with respect to the subject of his illness, such that even if he had been asked to speak on another topic, he would have been likely to revert to this matter fairly quickly. In contrast, Speaker 9, who had experienced a cerebellar stroke 7 years prior to the study, seemed reluctant to answer the preliminary interview questions about his health. When it became apparent that he was making reading errors, and the author tried to probe him on whether he had any visual difficulties, he was not forthcoming. This suggests that he may have been uncomfortable delivering a monologue on his health

status – a common choice when fixing the discourse topic in a disease population. In contrast to these challenges, it was thought that by inviting participants to speak on a topic of their choice, the outcome would be a monologue that is as natural as possible and is a faithful representation of the individual's level of functional communication in an informal environment. Despite these arguments, further research would certainly be worthwhile to assess the effect of different strategies for eliciting spontaneous speech. A final point worth raising in this context is that the aforementioned recommendation of assessing *multiple* monologues for a given speaker would help to reduce the confounding effect of subject matter, as it would enable calculation of an intelligibility measure that has been averaged over monologues with different levels of familiarity and linguistic complexity. The fact that Speaker 9 yielded a difference in SSI of 12% for two monologues that were supposed to be on the same topic illustrates that fixing the topic alone is unlikely to be sufficient to deal with the issue of linguistic and phonetic variability.

The study used a heterogeneous listening sample, ranging from individuals who had no formal experience of listening to disordered speech to experienced SLTs and phoneticians. This decision was made for logistic reasons, due to the lack of availability of sufficient numbers of listeners of one type. As explained in Chapter 2 (Section 2.4), the heterogeneity of the sample was a strong motivation for choosing a method of measuring SSI that was based on transcription (as opposed to a rating method). As discussed in the previous subsection, there have been mixed findings on the effect of listener experience on the orthographic transcription of dysarthric speech. However, on balance, the literature appears to indicate that experienced SLTs are able to decipher a greater proportion of the uttered words than naïve observers, particularly in the case of severe speakers. Furthermore, the present task differs from that used in previous studies, as (a) the stimulus consists of spontaneous speech rather than a reading task and (b) in addition to providing a transcription, the listener is required to count syllables in the unintelligible portions. These additional challenges are likely to enhance the difference in performance between naïve and experienced listeners. Due to the modifications made to Lagerberg et al.'s (2014) method in the present study, it was not possible to undertake any formal analysis to assess whether the transcriptions of experts were, on average, statistically different from those of lay listeners. However, as explained in Chapter 3, *all* perceptual data in this thesis underwent an outlier-removal process to identify responses that appeared to be markedly different from those of other listeners. In the case of spontaneous-speech analysis, a number of transcriptions from lay observers were removed due to the fact that the listener showed inferior perceptual skills. This was often a consequence of low familiarity with the Antwerp accent (a fact that came to light in response to the demographic questions asked

in Qualtrics). More interestingly, there was also one listener, an experienced SLT who works solely with adults, whose data were discarded because she consistently transcribed credible utterances that no other listener had perceived.[16] This suggests that differences in performance for the two listener groups (expert and non-expert) are likely, and that future studies should aim to measure intra- and inter-rater reliabilities for each group separately. This would allow determination of the feasibility of implementing the technique both as a research tool with naïve listeners and in clinical practice.

A further limitation of the study was the relatively low number of listeners. For the more severe speakers, in particular, there was considerable variability in the extent to which listeners were able to decipher the less intelligible portions of speech. This is illustrated in Table 7.1, where two out of five listeners perceived the words "*gewonnen heeft*" ("has won"). For utterances such as these which are of borderline intelligibility, the consensus of 8-10 listeners would have been preferable. The number of listeners was also insufficient to enable the spontaneous speech of the control group to be analysed. Due to the fact that some of these speakers had strong accents and used dialect words, the author was not able to determine whether their monologues would have been 100% intelligible to a listener familiar with the Antwerp accent. Therefore, future research is required to assess whether neurotypical speakers are at ceiling, and if not, to determine the cutoff for dysarthria diagnosis using the current SSI metric. Note, however, that the lack of normative data does not affect the specific research questions investigated in the present study.

7.5. Summary

The main goal of this study was to determine the degree of correlation between measures of intelligibility derived from single-word reading and a syllable-accuracy metric derived from unconstrained, spontaneous speech (Lagerberg et al., 2014). The level of correlation obtained when using C1 accuracy as the single-word intelligibility metric (r = 0.76, one-tailed $p < 0.01$) is in line with that reported in previous studies. It reinforces the premise of this thesis, namely that errors identified by phonetic-contrast analysis are predictive of real-world intelligibility. However, since correlation does not imply causation, further work is required to determine whether articulatory therapy results in a commensurate improvement in intelligibility in everyday speech. The second objective was to assess the suitability of Lagerberg et al.'s metric for measuring spontaneous-speech intelligibility in adults with dysarthria. Due to the modifications made to Lagerberg et al.'s method in the

[16] The SLT was not personally familiar with the speaker, who was a patient at a different hospital. The participant in question was S8, the gentleman with cognitive-communication difficulties.

present study, a rigorous assessment was not possible, and this is left as a matter for future research. However, it is possible to draw some preliminary insights. The first component of the Lagerberg method is to transcribe the intelligible portions of speech. The major threat to the validity of this component is that the decision as to whether to label a particular utterance as "intelligible" is left in the hands of the listener. If future research demonstrates that listeners are unreliable in making this assessment, then possible solutions might be to acquire a transcript of the monologue from the speaker (i.e., a gold standard) or to implement a consensus approach (assuming that data from multiple listeners can be acquired). The second component of the method is to count the number of syllables in the unintelligible portions. In the author's experience, this is not a straightforward task, and an accurate estimate may require more time and attention than could reasonably be expected from listeners in a research study. It could also be infeasible for implementation of the technique in the clinic. Again, a potential solution would be to obtain a transcript of the monologue. Alternatively, it may be possible to devise automated methods of deriving syllable counts that are sufficiently accurate to produce SSI measures of high reliability and validity. The extent to which these sorts of solutions would be feasible and worthwhile, either in a research context or in clinical practice, would depend on the specific circumstances, including the reasons for assessing spontaneous-speech intelligibility.

8. General discussion

This discussion is separated into the two types of contribution to knowledge made by this thesis. The main goal of the thesis was to address a set of *methodological* questions. The findings with respect to these questions are summarised in Section 8.1, while the broader methodological implications for dysarthria assessment are discussed in Section 8.2. The secondary objective of the thesis was to gain preliminary information about *articulatory errors in Belgian Dutch dysarthria*. The findings on this matter are discussed in Section 8.3. The final two sections present suggestions for future work (Section 8.4) and a summary of the main conclusions of the thesis (Section 8.5).

8.1. Summary of methodological findings

The starting point for this thesis was the assumption that articulatory errors play an important role in real-world intelligibility, and that the perceptual identification of such errors, along with their categorisation according to some type of theoretical framework, would be a worthwhile endeavour for many speakers with dysarthria. Following a review of the literature, it was identified that Kent et al.'s (1989) method of phonetic-contrast analysis showed promise for the identification and categorisation of segmental speech errors, largely because the errors detected by such a method are inextricably linked with word *meaning*. However, the applicability of phonetic-contrast analysis to languages other than English had not been widely investigated. Moreover, most of the underlying assumptions of the technique had not been rigorously tested even for speakers of English. In particular, the following list of methodological research questions was identified:

1. Is the range of phonemic-substitution errors typically observed in Antwerp Dutch speakers with dysarthria adequately represented by a reasonable number of phonetic-contrast categories?
2. Is there reasonable agreement between the phonetic-contrast error profiles identified by (a) different listeners and (b) the same listener on different listening occasions?
3. What is the threshold for detecting dysarthria using single-word intelligibility testing?
4. Are there phonetic-contrast categories that should be excluded from a clinical assessment of Belgian Dutch dysarthria because the error rates in speakers with dysarthria are not significantly different from those seen in neurotypical speakers?
5. Are there significant differences between the word-accuracy scores and phonetic-contrast error profiles yielded by an open and a closed listener response format?
6. Are the number and types of error identified by phonetic-contrast analysis predictive of real-world intelligibility?

A series of listening studies was carried out that aimed to address these questions to the extent that was possible given the available time frame and resources of the project.

Chapter 4 presented the results obtained from orthographic transcription of a single-word reading task carried out by speakers with dysarthria from the Antwerp region of Belgium (n = 10). The word list was developed by the author and had not previously been tested. The main purpose of the study was to address Question 1. An analysis of inter-rater agreement (Question 2a) was also performed, but the methodology was suboptimal (due to the limited number of listeners), meaning that the findings should be regarded as preliminary. The study revealed that phonetic-contrast analysis shows considerable promise with regard to consonant confusions: at least 78% of the substitutions observed in a given speaker could be coded using 13 phonetic-contrast categories. For vowels, on the other hand, many of the observed confusions did not lend themselves to categorisation based on reasonably well-defined phonetic features (e.g., vowel height or backness). The study further revealed that, on average, 39% of a speaker's phonetic-contrast errors on any given target word were unique (i.e., only heard by one listener). If future work, using a sample of expert listeners, were to determine that the intra- and inter-rater reliabilities are low for the *outcome measure* (the profile of phonetic-contrast errors), then the technique may not be suitable for clinical use, unless combined with instrumental analysis.

The purpose of Chapter 5 was to acquire normative data (n = 8) that would provide answers to Questions 3 and 4. Turning our attention firstly to Question 4, a number of phonetic contrasts failed to show evidence of being more vulnerable in speakers with dysarthria: the voicing of word-initial stops, nasal place confusions (at both word positions), and the vowel substitutions /ɛ/ → /ɪ/ and /ɪ/ → /i/. There may be other contrasts that are equally vulnerable in neurotypical speakers (e.g., /h/ deletion), but could not be detected in the present study due to low statistical power. Having identified confusions that are unlikely to be indicative of dysarthria, the errors observed for these categories were removed from the analysis, and the remaining normative data were used to calculate threshold word-accuracy scores below which an individual would receive a diagnosis of dysarthria. The cutoffs for the 95% and 97.5% confidence levels were 88.5% and 87.5%, respectively.

Chapter 6 aimed to determine the effect of the response format (free vs. forced choice) on word-accuracy values and phonetic-contrast error profiles in speakers with dysarthria (Question 5). Word accuracy was found to be significantly higher in the forced-response mode. The absolute difference in the percentage of correct words ranged from 4.3% to 24.6% across the cohort (n = 8), with a mean (± 1 SD) of 13.1% ± 6.9%. It was reasoned that this difference can be largely attributed to the fact that the free-response mode

increases the likelihood that a phonetic distortion will be perceived as a phonemic substitution. However, it is also possible that some substitution errors go undetected in the forced-choice mode due to bias. The speakers were not ranked in precisely the same order in the two response modes; however, the two sets of word-accuracy scores followed a similar trend (Pearson's r = 0.86, one-tailed p = 0.003). Fleiss' kappa was calculated to determine the level of agreement on responses to individual test items. There was moderate agreement for six out of eight speakers, with fair agreement and slight agreement for the remaining two speakers respectively. As was also mentioned for the free-response study, further work is needed to determine intra- and inter-rater reliabilities for the actual outcome measure (the profile of phonetic-contrast errors). The last analysis in the study compared error profiles for the two response modes. For both vowel and consonant contrasts, the response modes showed differences in the top six error categories, which were defined on the basis of error-ranks calculated for each speaker and then summed over the cohort. For consonant contrasts, the correlation between the summed ranked errors for the two response modes was r = 0.735 (one-tailed $p < 0.001$). The corresponding value for vowels was 0.622, p = 0.006. The correlation values were often substantially lower when calculated for individual speakers, particularly in the case of vowels. It is likely that the weaker correlations for vowels were largely due to the fact, unlike consonant categories, vowel categories were defined as substitutions between specific phonemes.

The last study in the thesis (Chapter 7) addressed Question 6 by examining the degree of correlation between single-word reading (SWR) intelligibility and spontaneous-speech intelligibility (SSI) in speakers with dysarthria (n = 10). The SSI metric, which was based on the method proposed by Lagerberg et al. (2014), was the ratio of the number of syllables perceived in the intelligible portions of speech to the total number of syllables perceived in the monologue. Reasonable correlation (r = 0.76, one-tailed $p < 0.01$) was obtained when using initial-consonant accuracy as the measure of SWR intelligibility. However, there were individual speakers who seemed to depart from the trend, the most striking case being that of a lady who was much more intelligible in spontaneous speech than would be expected based on her segmental accuracy score. Furthermore, perceptual assessment of her spontaneous speech revealed that the number of clearly identifiable phonemic-substitution errors was low. This finding was interpreted as further evidence to support the contention that many of the substitution errors perceived using orthographic transcription are in fact more likely to be distortions. The study also contributed knowledge on the challenges associated with using orthographic transcription to calculate a measure of spontaneous-speech intelligibility in scenarios where the intended output of the speaker is unknown.

8.2. Methodological implications for dysarthria assessment

8.2.1. Free versus forced choice

The finding that open and closed response modes provide information that is qualitatively different is not surprising. In fact, one of the causes of the difference was known in advance, namely that orthographic transcription identifies *prominent* phonetic confusions, some of which will be a consequence of the contrast in question having a high functional load, while the forced-choice mode identifies *vulnerable* confusions. However, it was argued in Chapter 6 (Section 6.4.2) that differences arising from functional load considerations are unlikely to present a problem in the long term, as the acquisition of a large amount of orthographic-transcription data from speakers with and without dysarthria would enable its confounding effect to be understood and corrected for.

Of greater importance are the other two mechanisms that underlie the qualitative differences between the two response modes. Firstly, it was argued that distortion errors are more likely to be heard as phonemic substitutions when the listener's response is unconstrained. In Chapter 7, it was argued that in this respect, the multiple-choice mode may have greater functional relevance than orthographic transcription. This is because distortion errors are also less likely to be a barrier to intelligibility in spontaneous speech, where the listener can make use of additional cues and context. However, there is insufficient evidence at present to state with certainty that distortion errors have no functional relevance. Furthermore, it could be the case that only distortion errors (and not substitution errors) are under the speaker's control and hence amenable to therapy.[1]

It was hypothesised that the second mechanism by which errors "disappeared" in the forced-choice mode was due to bias. A source of bias that can definitely be regarded as undesirable is when a clear substitution error is produced, but the error in question is not presented to the listener as one of the distractors. This is particularly likely to occur if the error belongs to a category that is not phoneme-specific, such as 'final singleton → cluster'. The ability to detect syllable-shape errors is especially important for languages (such as Dutch) that contain large numbers of complex consonant clusters. Yet the forced-choice mode is unlikely to be able to detect all instances of such errors, even if the number of distractors is substantially increased beyond that used in the present study. Another scenario in which bias may mask a genuine substitution error is when the speaker consistently produces substitution errors on multiple word segments. If listeners have a

[1] Even if the correction of distortion errors does not increase intelligibility, it could improve the *naturalness* of spontaneous speech.

propensity to choose one type of error over another (e.g., consonant errors over vowel errors), then this could result in the less salient error going undetected.

In summary, the open mode is likely to produce a greater number of "false-positive" findings than the closed mode – that is, phonetic distortions classified as phonemic substitutions. On the other hand, the forced-choice mode has the potential to yield false-negative outcomes, because inherent sources of bias could cause some of the speaker's phonemic substitutions to go undetected. Further research is required to understand the extent of these two limitations, particularly in the case of the forced-choice mode, where there may be potential for reducing the amount of bias, at least in some dysarthria assessments – for example, by increasing the number of distractors. The decision as to which response mode is more appropriate for any given scenario is likely to rest on factors such as the severity level of the speaker and the goals of the assessment. For example, for speakers of higher intelligibility who are being considered for articulatory therapy, the free-response mode might be more appropriate, as it could increase the likelihood that at least some of the individual's articulatory deficits will be identified.

8.2.2. Characteristics of the single-word stimuli

There are a large number of stimulus characteristics that are likely to affect the outcome of a single-word intelligibility assessment. Focusing on those that are most relevant to the current set of stimuli (monosyllabic, real words that are highly contrastive), the list might include: word frequency, word imageability, phonetic composition (both of individual words and of the word list as a whole) and sound-orthography correspondence. The order in which the speaker utters the words could also be important, due to factors such as fatigue, task accommodation and priming.

One of the most important design considerations concerns the trade-off between a phonemically-balanced word-list and a set of stimuli that tests most or all of the phonemes of the language with approximately equal frequency. An approximately phonemically-balanced word-list was used in the present study (85%), on the grounds that it would maximise the functional relevance of the error profiles and intelligibility scores. However, it was sometimes the case that very few errors were perceived on the more frequent phonemes of Dutch, while errors on less frequent phonemes (which are often those that are thought to be more difficult to produce) arose more often. Consider Speaker 7, for example. Her word accuracy in orthographic transcription (88.4%) was higher than that of three of the neurotypical speakers. However, she yielded more errors on /v/ and on word-initial clusters than any of the control subjects. Thus the advantage of including difficult

contrasts more often than they naturally appear in the language is that errors on these contrasts might be markers for dysarthria. Therefore, the optimal approach might be to aim for a compromise in which the word list is approximately phonemically-balanced, but phonemes that are known to be vulnerable in dysarthria are tested on a sufficient number of occasions such that the error metric has reasonable reliability.

Some preliminary evidence was gained in this thesis regarding the importance of lexical frequency. It was shown that, on average, there was no significant correlation between word frequency and error rate. However, inspection of the data revealed that words of low frequency should be avoided, as they may result in atypical errors that are not necessarily informative about disordered speech (as evidenced by the fact that the error is also seen in neurotypical speakers and only for the target word in question). An example in the present study was the target word /ʃɔu/, meaning '*(I) haul*', perceived as /ʃa:l/, meaning '*scarf*'.

Other factors that may be influential, such as phonetic context, were not examined in the present study, as a thorough analysis of the effects of such variables would require a larger sample size. As was argued with respect to the phonemic distribution of the sample, there may be a trade-off between setting these parameters to be representative of everyday language and using stimuli that are sensitive to dysarthria, so as not to "waste" test items. Bunton and Weismer (2001) reported that of the 13 word pairs used to test the high-low vowel contrast in Kent et al.'s (1989) assessment, four of these word pairs did not produce a single error in any of their speakers (25 dysarthric and 10 normal controls). Inspection of Kent's word list reveals that three out of these four vowel contrasts (*knew* → *gnaw*, *geese* → *guess* and *had* → *hid*) were unique in the sense that they involved phonemic substitutions that were not tested by any other items. Furthermore, of these three unique contrasts, two of them were pitted against another vowel-height distractor for the same target word (i.e., *gnaw* competed with *know* and *guess* competed with *gas*). Therefore, the fact that these three word pairs did not yield any errors in Bunton and Weismer's (2001) study is unsurprising. In the present study, there were also categories for which particular target words did not produce any errors (in either response mode). However, in most cases, there was no obvious reason as to why the word might be resistant to errors. Therefore, prior consideration of factors such as lexical frequency and phonetic context may not be sufficient to lead to an assessment that is optimal in the sense that there are no targets (or word pairs in the case of a multiple-choice assessment) that are immune to errors. This suggests that optimisation of the word list is likely to require the collection of a large amount of test data, from both speakers with dysarthria and neurotypical controls. Finally, as noted by Miller (2013), if *parallel* word lists are required (to avoid the problem of

listener familiarity), then these should be matched as closely as possible on the linguistic and phonetic characteristics of the stimuli.

8.2.3. Elicitation mode

There are a number of options for eliciting single words from speakers, including word reading, word repetition and picture naming. Picture naming has the obvious disadvantage that it places an extra constraint on the choice of stimuli. Word repetition is not representative of real-world communication and it may suppress errors, as some speakers might benefit from hearing an accurate exemplar of the target. However, word reading is not without its drawbacks either, as it may be unsuitable for people with visual difficulties or with aphasia that predominantly affects reading. Therefore, it would be useful to determine whether picture naming is a viable alternative for such individuals, and whether it produces similar intelligibility scores and error profiles to word reading. It was hypothesised in Chapter 3 (Section 3.4.2) that one might expect greater accuracy for word reading than for picture naming, due to the presence of orthographic cues in the former case. A preliminary investigation was carried out to test this hypothesis for two speakers in this study, S5 and S9. Their picture-naming stimuli were orthographically transcribed by three and five listeners, respectively. Word accuracy was compared with the accuracy for the same set of words in the reading task. Speaker 5 yielded an increase in word accuracy of 13.9% between word reading and picture naming (83.3% → 97.2%), while Speaker 9 exhibited an increase of 7.4% (68.4% → 75.8%). Therefore, according to this preliminary investigation, orthographic cues did not seem to enhance articulatory precision. If future studies were to confirm that picture naming yields higher intelligibility than word reading, a possible explanation could be related to the amount of articulatory effort. It was noticed during the interviews that word reading seemed to be carried out without much self-awareness or engagement on the part of the speaker. In contrast, the picture-naming task seemed to spark the speakers' interests, and on some occasions, when they believed that they had identified the correct word, it was uttered in the style of a "eureka moment". These observations would be consistent with an explanation based on articulatory effort. Substantially higher intelligibility scores for picture naming would suggest that the technique may not be suitable for mild speakers. In severe speakers, on the other hand, the technique may help to distinguish between articulatory errors that are and are not under speaker control. Such information could be useful for planning intervention.

8.2.4. Error categorisation

One of the main goals of this thesis was to assess the feasibility of categorising dysarthric errors in terms of phonetic contrasts. Such a framework enables the identification of speech deficits of a more general nature than, say, a list of common phonemic-substitution errors. Thus, the assessor is saved the effort of interpreting information on a very detailed level (e.g., a confusion matrix). Furthermore, the findings provide immediate information about the nature of the impairment for a given speaker or neurological group. For example, Kent et al. (1990) showed that phonatory and velopharyngeal functioning were the two most affected subsystems in male speakers with ALS. A further advantage of describing errors in terms of phonetic contrasts is that it reduces the number of test stimuli, as it can be assumed that all errors of a specific type (e.g., 'fricative place') are predicated on the same mechanism.

The present thesis identified a number of limitations associated with Kent et al.'s (1989) framework, at least in Belgian Dutch speakers. Firstly, with the exception of the durational contrast represented by /a:/ - /ɑ/, it was not possible to categorise monophthongal vowel confusions in terms of a change in (predominantly) one phonetic feature. Some of the vowel substitutions that involved only a small shift in the vowel space could be described as mainly a vowel-height contrast, such as /u/ → /o:/, /ɛ/ → /ɪ/ and /ɪ/ → /i/. However, such confusions were also common in control speakers, and it is not clear that they are indicative of a production impairment. The vowel confusions that were more clearly "dysarthric" either involved an approximately equal shift in backness and height, such as /ɑ/ - /ɛ/ and /ɔ/ → /ɑ/, or they yielded very low error rates in the current cohort (e.g., /ɔ/ - /u/). From inspection of Fig. 4.6, the situation can best be summarised by stating that contrasts in backness did not occur without a simultaneous contrast in height. It is possible that a similar observation was made by Haley et al. (2000) in their orthographic-transcription study of American English speakers, as they stipulated that vowel confusions involving both frontness and height should be coded as a front-back confusion only. However, the strategy of separating vowel confusions into these two categories (i.e., height only and height + backness) did not seem to be applicable to the present study because, as mentioned, the vowel confusions that predominantly involved a shift in height were either not particularly common or not "dysarthric". Further research would be required to determine whether this finding holds for a larger sample size and whether it is peculiar to Antwerpian Dutch. Certainly, there are word pairs in Kent et al.'s assessment that, in General American English, are predominantly a shift in vowel height and that involve a reasonably large shift across the vowel space (e.g., *him* - *ham*, *shoot* - *shot*). Furthermore,

Bunton and Weismer (2001) reported that errors were observed for these word pairs in a multiple-choice assessment with 25 dysarthric speakers. Therefore, from these observations, there is some evidence to suggest that the categorisation of vowel errors using phonetic features may be more challenging for Antwerpian Dutch than for English, due to the configuration of the vowel space. However, further research is required. As mentioned in Section 8.2.2, there may be scope for optimising the current word list, by replacing test stimuli that are resistant to errors. This process could increase the likelihood of perceiving "dysarthric" vowel-height errors, e.g., /ɔ/ - /u/. If this is not the case, then a different approach to categorising vowel errors in Belgian Dutch speakers may be needed.

A second limitation of the Kent et al. (1989) framework was that errors for a particular contrast category were often either unidirectional or involved only a subset of the phonemes to which the contrast pertained. For example, for the category 'fricative place', errors involving the alveolar fricatives were rare, and fricative backing was far more common than fricative fronting. The 'stop vs. fricative' category was also found to be (approximately) unidirectional, but at the level of the individual speaker. In other words, a given speaker tended to yield errors in one direction only: either the stopping of fricatives or the frication of stops. This suggests that these two errors result from a different type of motor deficit. Therefore it seems that whether the goal is to identify targets for therapy or to learn more about the nature of the impairment, contrast errors should at least be broken down according to the error *direction*. In the case of planning intervention, it may also be necessary to examine the specific phonemes involved. Otherwise, goals could be set relating to phonemes that are not prone to error. Thus, there appears to be a trade-off between obtaining information that is highly relevant and gaining a broader perspective of the nature of the impairment. A possible solution would be to carry out an initial assessment based on the full set of contrast categories and phonemes, followed by a more focused assessment to obtain reliable information about the phonemes and phonetic contrasts that are most vulnerable. A similar suggestion was made in a more general context by Miller (2013). A two-tier process could be particularly beneficial for speakers with mild dysarthria whose errors are often confined to just a few phonetic features.

Finally, there was evidence in this thesis to suggest that it may be important to distinguish between true substitutions (misarticulations that cross a phoneme boundary) and productions that are perceived as substitutions despite the fact that they are more likely to be distortions. The ratio of distortions to substitutions was hypothesised to vary among speakers, even when they yielded a similar word-accuracy score in orthographic transcription. For example, there was evidence to suggest that S4 produced fewer

substitutions than S9, including the fact that she was more intelligible in spontaneous speech (86.5% vs. 73.0%). Yet, in the free-response mode, Speaker 4 yielded a lower word accuracy (49.1% vs. 62.0%). Given that substitutions seem to be more disruptive to real-world intelligibility than distortions, it could be worthwhile developing a method to ensure that only true substitution errors (or at least major distortions) contribute to the phonetic-contrast error profile. In word-recognition studies that involve multiple listeners, it could be the case that this information arises naturally; i.e., perhaps the greater the number of listeners who perceive the error in question, the more likely it is to be a substitution rather than a distortion. When only one assessor is available, the use of a *closed* response mode might be the most efficient method of guarding against 'false positives', as argued in Section 8.2.1. A further possibility would be to produce, in the first instance, a contrast-error profile based on single-word recognition, and then to follow this up with a perceptual assessment of spontaneous speech to determine which of the contrast errors are still perceptible.

8.3. Articulatory errors in Belgian Dutch dysarthria

The secondary objective of this thesis was to obtain information about phonemic and phonetic-contrast errors observed in Belgian Dutch speakers with dysarthria. Detailed discussion of the vulnerability of specific *phonemes* was provided in Chapters 4 and 5, so this issue will not be considered further here. Regarding *phonetic-contrast* errors, it was argued that a full understanding of these errors would require integration of information from Chapters 4-6, to deal with the two main confounding factors: perceptual similarity and functional load. Such an analysis was carried out at the end of Chapter 6, resulting in two tables (one for vowels and one for consonants) listing the directional contrast categories deemed to be "important" in the present cohort. The reader is reminded that the criteria for an important error were that it should be (a) frequently occurring (*prominent*) in the free-response mode, (b) *dysarthric* and (c) *vulnerable* in the forced-choice mode. The remainder of this section discusses the implications of the findings in Tables 6.7 and 6.8 with regard to motor speech impairment in dysarthria.

A large proportion of the segmental errors perceived in this thesis arose in participants with cerebellar injury or disease. As shown in Table 4.1, this was the site of damage for two out of the four least intelligible speakers (S5 and S9) and well as for two further speakers: S3 and S6. The least intelligible speaker, S4, had sustained surgical damage following resection of a hemangioblastoma in the fourth ventricle of the posterior fossa, and was judged by neurologists to exhibit signs of ataxia in clinical examinations (e.g., a slow, unstable gait, inability to walk in a straight line, slow right-sided dysdiadochokinesia, and

ataxic right-sided heel-knee-shin test). A T2-weighted MRI showed residual hyperintensity in the posterior cerebellum. Note, however, that this speaker also had right-sided tongue paralysis, probably due to surgical damage of the left-sided hypoglossal nerve, which would have been a major cause of her dysarthria.

Before embarking on a discussion of "important" Belgian Dutch dysarthric errors, a brief summary is provided on *general* aspects of speech production in cerebellar dysarthria. Previous findings that relate to *specific* deficits are mentioned in the subsections that follow. The cerebellum is the part of the brain thought to be most involved in the coordination and sequencing of movements. Kent et al. (2000b) defined *sequencing* as "the order of succession, as in the case of phonetic segments, movements, or muscle contractions". They defined *coordination* as "the processes of adjustment by which separate components of action are unified in a common motor objective". Given the generic role of the cerebellum in executing movement, it is unsurprising that the perceptual and acoustic features of cerebellar dysarthria reflect global impairment of the respiratory, laryngeal and articulatory subsystems of speech (Kent et al., 2000a). However, the impairments are thought to be less consistent than in other dysarthria types, as captured by the Darley label "irregular articulatory breakdowns". For example, vocal pitch or loudness may suddenly change from one moment to the next, and consonant articulation may switch between lenis and fortis productions (Ziegler, 2016). In connected speech, altered stress patterns are often reported, including the disturbance known as excess and equal stress. An assessment of syllable alternating motion rate typically shows a slow and irregular temporal pattern. Physiological measurements reveal an overall slowing of muscle movement, as well as bursts of excitation and quieting (Kent et al., 2000a). A small number of instrumental studies have shown respiratory dysfunction, including irregularities in breathing patterns during sustained vowel phonation and syllable repetition (Kent et al., 2000a).

In the following subsections, unless stated otherwise, when the number of instances of a particular error is mentioned, this refers to the *free-response* study. This is because there were far fewer errors in the forced-choice mode, making comparisons among speakers more difficult. On occasions where the discussion requires understanding of the role played by functional load, errors from the forced-choice study are instead reported. A further point to note is that a contrast confusion was only considered "important" in Tables 6.7 and 6.8 if it was both prominent and vulnerable. Since the purpose of the following discussion is to draw insights about impaired speech production, it could also be interesting to include categories that were vulnerable in the forced-choice study but not prominent in the free-response mode (due to low functional load). However, there were only two confusions that

met this criterion (initial /h/ addition and the diphthong confusion /œy/ → /ɔu/), and in both cases, the number of potential errors in the forced-choice study was small. Therefore, the finding that these contrasts were "vulnerable" may be anomalous. Nevertheless, the glottal vs. null contrast is briefly discussed in the first subsection below (on syllable-shape confusions), as it is a category that has produced high error rates in previous populations.

8.3.1. Syllable-shape confusions

A prominent error perceived in the orthographic-transcription study was phoneme addition. In particular, at word-final position, there were 109 instances of the perception of a cluster instead of a singleton, but only 14 instances of 'final-cluster reduction'. There was some evidence to suggest that final singleton → cluster confusions may be particularly prominent in individuals with cerebellar damage or disease. Speaker 5 yielded 23 of the 109 instances of this confusion, while S4 was responsible for 61 instances (although note that this speaker had right-sided tongue paralysis in addition to her ataxic symptoms). Typical examples of the confusion (e.g., /kɪn/ → /kɪnt/ and /ko:rt/ → /ko:rts/) reveal that the tendency was for listeners to perceive the addition of a homorganic phoneme produced in a different manner. The other word-shape category that yielded a relatively large numbers of errors was 'initial singleton vs. cluster', although in this case, the errors were more evenly distributed over the two directions: 55 instances of initial singleton → cluster and 35 instances of initial-cluster reduction. A difference of this order of magnitude could be due to the effect of functional load. The gentleman with a cortical watershed stroke, S8, who was the third least intelligible speaker in the cohort (based on single-word reading accuracy), was responsible for 23 of the 35 instances of initial-cluster reduction. Confusions in the opposite direction (initial singleton → cluster) were distributed over nine speakers and did not show any obvious pattern with regard to site of neurological damage.

Further research is required to determine whether the final singleton → cluster confusion is a hallmark of cerebellar dysarthria. However, it is worth noting that the error is consistent with a deficit characterised by reduced speed, a lack of coordination and irregular timing. In the case of word-final /n/ → /nt/ confusions, for example, it seems that the speaker raises their velum before releasing the occlusion. For /t/ → /ts/, it is possible that rather than releasing the stop abruptly in one clean gesture, the speaker leaves their tongue in sufficient proximity to the alveolar ridge to produce turbulence. To the best of the author's knowledge, distortions of this type have not previously been reported as a feature of cerebellar dysarthria. However, they have been mentioned in the context of apraxia of speech (Kent & Rosenbek, 1983), where the authors noted that segmental speech errors in people with AOS were indicative of a "disorder in the selection, retrieval, or

seriation of phonemic or phonetic units." The listed examples included /twɛlvz/ for English *twelve* and /drɪʃəz/ for *dishes*. Thus, although deficits of this nature would occur at a different stage of speech production in speakers with apraxia and speakers with dysarthria, in both cases, they would signify disordered sequencing.

Finally, it is worth mentioning the glottal vs. null category, which has produced high error rates in a number of previous studies involving different clinical populations (e.g., Kent et al., 1990;[2] Blaney & Hewlett, 2006; Whitehill & Ciocca, 2000b; Bunton & Weismer, 2001). In the present study, /h/-deletion was deemed to be an "important" error, but further research is required to accurately determine its incidence in neurotypical speakers from Antwerp. Confusions in the other direction (null → glottal fricative) were not tested sufficiently often in the present study to enable reliable conclusions to be drawn about vulnerability (see Chapter 6, Section 6.3.3). It is also difficult to draw conclusions about the vulnerability of /h/-addition from previous studies, as the predominant *direction* of phonetic-contrast errors was not usually stated. An exception is Blaney and Hewlett (2006), who reported that glottal vs. null errors were approximately equal in both directions (11% deletion vs. 9% addition) in a group of speakers with dysarthria due to Friedreich's ataxia. They argued that the prominence of 'glottal vs. null' errors was consistent with the high error rate observed in these speakers for the 'initial plosive voice' category, as both errors imply deficits in laryngeal timing and control. Based on these arguments, it seems likely that /h/-addition will prove to be a common error in Belgian Dutch dysarthria, at least in some clinical populations.

8.3.2. Voice confusions

The 'stop and fricative voice' category, which only applies to the word-initial position in Dutch, was both prominent in the free-response mode and vulnerable in the multiple-choice mode. In both response modes, *devoicing* was the predominant error direction. This finding (a higher incidence of devoicing) was driven by the directionality for *stops*, as these were responsible for most of the errors in the 'stop and fricative voice' category.[3] Based on the present findings, the *voicing* of stops was not considered to be a dysarthric error at all, as it was no more prominent in speakers with dysarthria than in neurotypical speakers.

[2] This study involved male speakers with ALS. In contrast, the glottal-null category produced very few errors in female speakers with ALS (Kent et al., 1992). The authors could find no obvious explanation for this difference.

[3] In fact, for fricatives, *voicing* was more common than devoicing in orthographic transcription. However, this could be due to the effect of functional load. It is not possible to draw any conclusions about directionality for the MC mode, as there were only two word-pairs that tested fricative voice.

Previous studies that implemented Kent et al.'s (1989) method[4] in speakers with different neurological conditions likewise reported that the 'initial voice' category is one of the most prone to error (see, for example, Kent et al., 1990; Blaney & Hewlett, 2007; Bunton & Weismer, 2001; and Gentil, 1992, who developed a French version of the Kent et al. test and applied it to speakers with Friedreich's ataxia). Therefore, the vulnerability of the word-initial voice contrast seems to be a feature of different languages and of dysarthria arising from different causes. In Dutch, the voiced plosives are realised with a negative voice onset time (pre-voicing) and are considered to be difficult to produce and relatively prone to disruption (see Chapter 5, Section 5.4.3). Therefore, it is not surprising that stop devoicing was found to be more common than stop voicing. Comparison of this finding with previous studies is complicated by the fact that the directionality of the perceived errors was not always reported. Furthermore, the voicing contrast is achieved by different means in different languages, and may involve a different set of phonemes; thus cross-linguistic agreement might not always be expected. Nevertheless, it is worth summarising some of the most relevant findings, which are those reported in studies that documented segmental, perceptual errors. Johns and Darley (1970), who investigated 10 American English speakers with different dysarthria types, observed just two word-initial voice substitutions, both of which were devoicing of /b/. Platt et al. (1980b) observed two instances of voicing of /t/ in 48 Australian male speakers with cerebral palsy. There were no errors for /b/ - /p/. The velar contrast, /k/ - /g/, which does not exist in Dutch, was the most prone to error, with three substitutions in each direction. Blaney and Hewlett (2007), who studied Irish speakers with Friedreich's ataxia (FA),[5] reported a higher incidence of word-initial voicing than devoicing (7% vs. 2%). Antolik and Fougeron (2013) investigated French speakers with Parkinson's disease (PD), ALS and cerebellar ataxia. Errors were identified using a combination of perceptual assessment and acoustic analysis, and instances of *partial* voicing / devoicing were included. The voice contrast was assessed using word-initial instances of /t, d, k, g/, which were extracted from a reading passage. There were approximately twice as many devoiced as voiced plosives, meaning that there was a higher *a priori* probability of voicing. The authors observed different patterns for the three groups. In ALS, voice distortions consisted almost entirely of voicing. In cerebellar ataxia, 29% of all distortions involved devoicing, compared to 24% for voicing. Devoicing was also the more common process in PD (37% vs. 16%). Thus it seems that the

[4] In studies that used Kent et al.'s assessment in its original form, the 'initial voice' category mainly referred to stops, as only one of the nine word-pairs used to test this category employs fricatives.

[5] This condition is thought to result in predominantly ataxic dysarthria mixed with spastic components (Folker et al., 2010).

directionality of 'stop voice' confusions is at least partly dependent on the neurological condition in question. In the present study, the largest neurological group (cerebellar: S3, S5, S6 and S9), in common with the rest of the cohort, showed a much higher incidence of devoicing. This is consistent with the finding of Antolik and Fougeron (2013) for the cerebellar ataxia group, as well as with other studies of French speakers with cerebellar dysarthria (Duez, 2014: p.178). It could be regarded as inconsistent with Blaney and Hewlett (2006), although note that FA dysarthria often includes a spastic component.

8.3.3. Place confusions

As shown in Table 6.7, the only place confusion that was found to be "important" in Belgian Dutch dysarthria was fricative backing. The fronting of fricatives was neither a prominent error in the free-response mode, nor did it yield high error rates in the closed response mode. Nasal place confusions were attributed to the low perceptual distinctiveness of this contrast, while oral stops yielded very few place of articulation errors in the MC mode. In fact, even fricative backing was not a widespread phenomenon in the sense that 11 of the 20 errors in the free-response study arose from just one speaker (S4). Furthermore, almost all of the 'fricative place' errors occurred for the labiodentals – a place of articulation that does not apply to stops. Therefore, it is possible that these errors reflect the vulnerability of labiodental phonemes (see Chapter 4, Section 4.4.3), rather than a consistent or major deficit in tongue positioning.

Consider, first of all, the speakers with a cerebellar condition. If future studies were to confirm that place of articulation does not appear to be a prominent error in cerebellar dysarthria, this would be consistent with electropalatography studies in speakers with spastic-ataxic and ataxic dysarthria (Folker et al., 2010), which showed that tongue placement during the articulation of lingual consonants is accurate.[6] As for other dysarthria types, previous studies reporting contrast-error profiles derived using Kent et al.'s (1989) method have often shown relatively low error rates for the place of articulation categories (with the exception of the alveolar-palatal fricative contrast), irrespective of the underlying neurological condition (see, for example, Kent et al., 1990; Bunton & Weismer, 2001; Whitehill & Ciocca, 2000b; Kent et al., 1992). This suggests that only *subtle* errors in place, involving relatively close positions, are observed in most dysarthric speakers. A similar finding was obtained by Kim et al. (2010) in their narrow-transcription study of speakers

[6] Atypical productions were instead characterised by a prolonged closure phase – thus a temporal deficit rather than a spatial one, which is in keeping with the notion that the cerebellum plays a role in timing. A prolonged closure phase for stops was also perceived (informally) in the present study. It was extremely prominent in S9, the gentleman with cerebellar dysarthria and scanning speech.

with cerebral palsy: place errors generally involved close positions such as labiodental vs. dental, dental vs. alveolar and post-alveolar vs. alveolar.

8.3.4. Manner confusions

The four error categories that involved manner confusions (stop-nasal, stop-fricative, /r/-/l/ and /r/-fricative) can all be regarded as strongly "dysarthric". In other words, manner errors were rarely perceived in neurotypical speakers, whereas voice errors were relatively common, and place errors arose now and again. Of these four categories, the two that involved *fricatives* produced the highest vulnerability rates (see Chapter 6, Fig. 6.5). This could be regarded as consistent with Platt et al. (1980b), who reported that fricatives were the most vulnerable consonant manner in speakers with dysarthria due to cerebral palsy. Furthermore, the acoustic signature of the feature 'fricative' was among the 15 most important predictors of phoneme intelligibility (assessed using the NSVO) in van Nuffelen et al.'s (2009b) phonological regression model.[7] The following paragraphs discuss each of the aforementioned manner categories in turn.

Stop-nasal confusions signify velopharyngeal incompetence. Table 6.7 shows that the stop → nasal contrast (nasalisation) was neither prominent nor vulnerable in the present study. In fact, it resulted in zero errors in the forced-choice mode. This finding may be peculiar to the current cohort, meaning that hypernasality was not a prominent feature of these particular speakers. On the other hand, it could signify that hypernasality in Belgian Dutch dysarthria does not easily manifest itself as a substitution error. Evidence against the latter argument is that 'stop vs. nasal' was the most affected category in Kent et al.'s (1990) study of male speakers with ALS. While the directionality of the errors was not stated, it seems highly likely that at least some of them were stops perceived as nasals (Eshghi et al., 2019). In the author's estimation, there are no obvious differences between English and Dutch that would lead to a much lower likelihood of this substitution being perceived in Dutch. As for the nasal → stop contrast, this was found to be an important category in the present study, although it only produced a vulnerability rate in excess of 0.1 for one speaker in the multiple-choice mode. The speaker in question, S9, was the gentleman with cerebellar stroke and scanning speech. Duffy (2005: p.175) states that intermittent hyponasality is perceived in some speakers with ataxic dysarthria, presumably due to "improper timing of velar and articulatory gestures for nasal consonants." Ziegler (2006) reports that irregular hypo- and hypernasality are both clinical features of ataxic dysarthria.

[7] As argued throughout the thesis, a high *correlation* with intelligibility does not necessarily imply a high average *vulnerability* for the feature in question. However, from visual inspection of published phonetic-contrast error profiles, it seems that in most cases, these two properties simultaneously apply.

For the stop-fricative category, errors in both directions were deemed to be important (see Table 6.7). The fricative → stop errors, which were mainly perceived at word-initial position, largely arose due to one speaker (S9), who yielded 18 of the 23 instances of this error in the open response mode. In all cases, the target sound was a labiodental, which, as mentioned, seems to be a challenging consonant-class. Fricative stopping is frequently observed in child speech and can be considered a simplification; fricatives are thought to be more difficult to produce than stops, as they require a greater degree of neuromuscular control. Confusions in the opposite direction were mainly confined to word-final position and were distributed across several speakers. The frication of stops can be regarded as "articulatory undershoot", and it has been suggested that it is a consequence of weakness and uncontrolled acceleration (Logemann & Fisher, 1981). A further observation in the present study was that a given speaker tended to either stop fricatives or fricate stops, rather than to yield errors in both directions. However, no clear pattern in directionality emerged across the cohort with respect to neurological condition. According to Ziegler (2016), both over- and undershoot are possible articulatory errors in ataxic dysarthria.

The /r/ vs. fricative category was both prominent and vulnerable in both directions. Confusions in the direction /r/ → fricative were mainly due to a single speaker (S4), who yielded 29 out of the 35 observations at word-final position and 16 out of 19 at word-initial position. The difficulty in producing an alveolar trill (the most common allophone of /r/ in the Antwerp accent) for this particular speaker is unsurprising, as she has right-sided tongue paralysis. As discussed in Chapter 4, the alveolar trill is considered a challenging phoneme to produce. In previous dysarthria studies (in other languages), this phoneme was observed to be de-rhotacised or substituted with /l/ or an obstruent. The confusion of /r/ with a fricative in the present study was hypothesised to be due to either a compensatory gesture on the part of the speaker or a misinterpretation of a distorted trill on the part of the listener (or indeed a combination of the two). Given that previous studies in other languages, which mainly used broad or narrow transcription (see Chapter 4, Section 4.4.3), reported different substitutions, it suggests that the confusion with a fricative in Dutch may be due to phonological factors, e.g., a high incidence of minimal pairs based on contrasts between /r/ and a fricative. Confusions in the opposite direction (fricative → /r/) mainly occurred at C1 position and were distributed across several speakers. The target sound was generally the velar fricative. Since the Dutch allophones of /r/ include uvular productions, the perception of /r/ upon hearing a misarticulated velar fricative is not surprising. In summary, there is some evidence to suggest that /r/ vs. fricative confusions may be particularly prominent in Antwerpian Dutch, and are explicable, at least in part, based on perceptual and phonological considerations.

Instrumental methods such as electropalatography would be a useful means of further investigating /r/ in Belgian Dutch dysarthria from a production perspective.

The final manner category that emerged as important in the present study was /r/ vs. /l/, with confusions in both directions labelled as prominent and vulnerable. In the case of /r/ → /l/, the perceived errors were confined to word-initial position and often occurred in consonant clusters. The most severe speaker, S4, was responsible for 15 out of 25 of these confusions. As discussed in Chapter 4 (Section 4.4.3), if a speaker is unable to produce the trill feature, a confusion with /l/ would be unsurprising. There was no evidence to suggest that the confusion was more common in the cerebellar group, and in fact, it was also observed on 8 occasions in one of the neurotypical speakers. Confusions in the opposite direction, /l/ →/r/, occurred at both word positions. The majority of these errors (10 out of 17) were due to S9, the speaker who showed excess and equal stress. Six of his errors occurred in word-initial consonant clusters. The listeners in the present study were, for the most part, proficient in English. Thus if a distorted /l/ sounded like the English alveolar approximant, it is not surprising that this would have been transcribed as "r". Substitutions between English /l/ and /r/ comprise one of Kent et al.'s (1989) contrast categories. However, in general, the category has yielded very few errors in previous studies with American English speakers (see, e.g., Blaney & Hewlett, 2007; Kent et al., 1990; Bunton & Weismer, 2001). An exception is Kent et al. (1992), who investigated female speakers with ALS. For two (out of the eight) speakers, the /r/-/l/ category was among the top 5 most vulnerable contrasts. However, the directionality of these errors was not reported. Further research with speakers of different languages would be useful for understanding this contrast, perhaps using instrumental analysis to investigate the nature of the articulatory deficit.

8.3.5. Vowel confusions

Table 6.8 listed the vowel confusions that were found to be important in the present cohort. Some of these errors are perhaps best described as reductions or simplifications, namely vowel shortening (/a:/ → /ɑ/) and monophthongisation. Both of these confusions are likely to arise in dysarthria from different aetiologies, and indeed, they were observed in every dysarthric speaker in the present study (as well in the control group). Diphthongisation was also perceived in every speaker with dysarthria, although it was much less common than monophthongisation. There was no evidence to suggest that the incidence of this error differed in speakers with and without cerebellar damage. According to Odell et al. (1991), diphthongisation of vowels could result from failure to make the necessary vocal tract constriction for adjacent consonants. Vowel *lengthening* was observed in half the cohort,

with 8 out of the 13 instances arising from the speaker who exhibited excess and equal stress in his spontaneous speech (S9). The lower incidence of vowel lengthening (compared to shortening) across the cohort (13 vs. 39 errors) seemed to be a genuine effect, as there was an approximately equal number of opportunities for each confusion to arise. Thus, the fact that S9 was the only speaker to yield a similar number of errors in each direction (8 lengthening vs. 7 shortening) is notable. He was also the only speaker for whom vowel lengthening persisted in the MC mode; for the remaining speakers who yielded lengthening errors in orthographic transcription, this category disappeared when listener responses were constrained. A correlation between scanning speech and inaccurate vowel duration would be logical, as both characteristics suggest difficulty with rhythm and timing. More specifically, if Speaker 9 produces all his syllables with approximately equal duration, which is the definition of scanning speech used by Ackermann and Hertrich (1994), then the unique pattern of vowel-duration errors seen in this speaker – an equal number of confusions in each direction – is exactly what one would expect.

The remaining confusions classified as important in Table 6.8 (shaded in grey) involve simultaneous contrasts in height and backness. Figure 8.1 plots the important monophthong confusions in terms of theoretical shifts across the F1-F2 vowel space for the average Antwerp speaker. In contrast to the vowel confusions reported in Chapter 4 (Fig. 4.6), Fig. 8.1 excludes confusions that were deemed to be "normal", as well as those that did not yield an appreciable error rate in the forced-choice study. Thus, it is assumed that only productions that were true substitutions (or at least major distortions) remain. Unlike previous figures of this type in the thesis, all contrasts are represented by a line of the same thickness. This is because the error rates calculated in the forced-choice study (which would be the metrics of interest in a discussion of impaired speech production) are based on low numbers of observations and are prone to undue influence from one or two speakers. Therefore, it would be imprudent to pay too close attention to the values obtained for specific confusions. Rather, the goal is to assess whether the pattern of errors reveals any consistent directionality in terms of height or backness.

As mentioned in Chapter 4, the degree to which vowel centralisation could ever be observed in a graph such as Fig. 8.1 is naturally limited by Antwerpian Dutch phonology, especially in a study where the word list was chosen to be phonemically-balanced. In particular, there is a scarcity of vowels in the centre of the vowel space, with the only truly central vowel (schwa) not being relevant to the present study and the next most central vowel (øː) occurring with relatively low frequency (meaning that it has a low prior probability of being perceived in a listening paradigm that is confined to real words).

Nevertheless, Fig. 8.1 does seem to reveal some evidence of compression of the vowel space in the front-back dimension; that is, for all the vectors that include a large horizontal component, the direction of the horizontal movement is towards the centre. As for vowel height, no consistent pattern emerges. The most striking observation is that the confusions that involve the two largest shifts in height, /ɔ/ → /ɑ/ and /ɛ/ → /ɑ/, involve increased *opening*. However, it is possible that these confusions were driven by the aforementioned tendency of speakers to centralise vowels in the horizontal direction. In other words, if speakers produce the vowels /ɔ/ and /ɛ/ with mid-range F2 values, then these tokens are likely to be perceived as /ɑ/, even if there is no accompanying vowel-height error, because there are no other possible confusions in the mid-F2 region. Of course, the converse could also be the case, i.e., that the main deficit was in vowel height and that the horizontal movement to the centre was due to phonological constraints. Based on the rest of the graph, this seems less likely; however, it should be borne in mind as a possible interpretation.

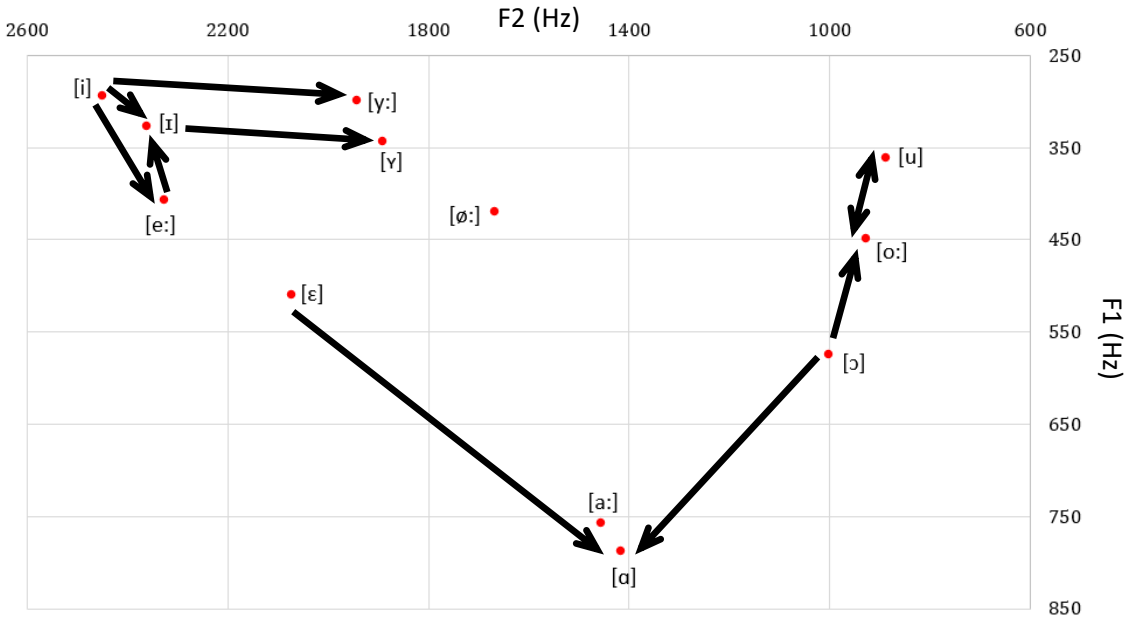


Figure 8.1. Monophthong confusions involving contrasts in vowel height and backness that were shown to be important in the present cohort.

The current findings should be regarded as preliminary, due to the low numbers of observations for some vowel contrasts and an incomplete understanding of the effect of bias in the multiple-choice mode. For example, /ɑ/ → /ɔ/ and /ɑ/ → /ɛ/ were two of the confusions that "disappeared" in the forced-choice mode, and while the interpretation favoured in this thesis is that these contrasts were distortions rather than substitutions, it is also possible that the errors disappeared due to bias (which could have had a differential effect on different vowel contrasts, even those produced with a similar level of phonetic distortion). Therefore, further research would be required to improve understanding of the

most vulnerable vowel contrasts in dysarthric speakers from the Antwerp region. Note, however, that when the goal of such research is to relate perceived vowel substitutions to impairments in jaw and tongue positioning, then a listener response mode that is based on the transcription of real words, or even on broad transcription, will be of limited usefulness. This is because, as mentioned, there is a paucity of vowels in certain regions of the vowel space, which places a substantial restriction on the range of substitutions that can be observed. Finally, it is worth mentioning that the phonetic characteristics of confusions between monophthongs and diphthongs were not examined in this thesis. It is possible that further categorisation of these confusions, in terms of changes in the dimensions of the component vowels, could provide information about overall deficits in jaw and/or tongue positioning.

8.3.6. Ataxic-dysarthria subtypes

Throughout the thesis, observations have been made about differences between the speech characteristics of S9 and those of the remainder of the cohort, including the other speakers with cerebellar injury or disease (S3, S5 and S6). Speaker 9 differed from the other cerebellar speakers in the following respects: (a) he showed a much higher incidence of fricative stopping; (b) he was the only speaker to yield /l/ → /r/ substitutions; and (c) he produced a large number of vowel duration errors, including vowel *lengthening* (a process not seen at all in S3, S5 or S6). Furthermore, Speaker 9 yielded the highest inter-listener agreement scores of the cohort and showed relatively little improvement in accuracy between the free- and forced-choice modes (compared to speakers of a similar severity level). In his monologues, he was the only participant who exhibited scanning speech and who seemed to produce consistent substitution errors across a number of different contrast categories. As argued in Chapters 6 and 7, many of these characteristics would be indicative of a speaker who tends to make consistent, well-formed substitution errors and who is fundamentally limited in his ability to produce certain sounds, irrespective of factors such as speech rate and level of effort.

It therefore seems reasonable to suggest that Speaker 9 may have a different type of dysarthria from the other cerebellar speakers – one that is characterised by greater consistency in his articulatory errors and prosodic excess. Duffy (2005: p.175-179) has also suggested that ataxic dysarthria may involve different subtypes. He surmises that there may be three classes of speaker: those with predominant prosodic excess, those with predominant articulatory inaccuracy, and a third group that has a "more equal combination of the two". Duffy intimates that the last of these three groups (to which Speaker 9 would belong) is the most paradoxical. This is because prosodic excess is a sign of *inflexibility* (a

lack of variability), while articulatory inaccuracy suggests the exact opposite – *instability* – especially in speakers with ataxic dysarthria, as they are known to exhibit "irregular" articulatory breakdowns. However, the articulatory errors of Speaker 9 appeared to be the most consistent of the entire cohort. This finding, if it could be replicated in other studies, would suggest a refinement to Duffy's subtypes hypothesis, with the three groups being defined as follows: (1) predominant prosodic excess, (2) predominant articulatory inaccuracy, characterised by *irregular* breakdowns, and (3) prosodic excess combined with *regular* articulatory breakdowns (i.e., both signs of inflexibility). The paradox raised by Duffy would then be resolved. Finally, it is worth noting in this context that the long-held view that ataxic dysarthria can be characterised by irregular articulatory breakdowns and excess and equal stress has been challenged by recent research, as reviewed by Mackenzie (2011).

8.4. Future work

The immediate next step would be to continue developing and testing the single-word reading assessment proposed in this thesis. A number of suggestions were made in Section 8.2 for optimising the word list and reducing the potential for bias in the multiple-choice mode. It could also be useful to devise a method of consolidating vowel errors into a smaller number of categories. The process of optimising the assessment should include analysing speech data from larger sample sizes with a wider variety of aetiological conditions. This would also allow for a comparison between male and female speakers, which could be important with respect to phonatory deficits (Kent et al., 2000a). Furthermore, additional normative data would be valuable for obtaining more accurate estimates of the vulnerabilities of (a) contrast categories that may not be dysarthric at all, such as /h/-deletion and certain vowel confusions (e.g., /u/ → /o:/), and (b) categories that have a dysarthric component, but are sufficiently vulnerable in neurotypical speakers such that cut-off error rates may need to be established in order to consider the contrast "disordered" (e.g., stop devoicing). Future normative studies should include speakers of a younger age than that investigated in the present study, so as to provide a point of comparison for younger people with acquired dysarthria. In addition, normative data for different age groups could improve our understanding of any phonological changes in progress among Antwerp speakers.

The underlying assumption of this thesis, which was tested in Chapter 7, was that articulatory errors exert an important influence on intelligibility in spontaneous speech. It was reported that the incidence of phonemic substitutions in the intelligible portions of

spontaneous speech was much lower than expected based on the single-word reading study. The explanation favoured in this thesis was that many of the errors transcribed in SWR using the free-response mode were in fact distortions, which became less perceptible in the context of a monologue. However, as discussed in Section 7.4.1, explanations based on speech *production* can also be imagined. For example, it was suggested that hyper-articulated, stressed syllables might exacerbate some segmental errors. One method of testing such hypotheses would be to compare error profiles for single words produced in isolation and for the same set of words embedded in connected speech. Patel et al. (2014) found that words uttered in isolation were just as intelligible as the same words produced in a sentence context. However, words that were extracted from the sentences and then presented in isolation had the lowest intelligibility. This line of research could be extended to determine the degree to which the articulatory precision of *specific contrasts* is reduced in connected speech. Data to enable such an investigation were acquired in the present study. The words in the word list were integrated into semantically implausible sentences such as /hɛt fɛin stɛl braːt də trɑp/ (*The great couple roasts the stairs*). These sentences introduce some of the factors that arise in spontaneous speech, such as syntactic cues and assimilatory processes. A further study could then be carried out to achieve closer verisimilitude to spontaneous speech. For example, participants could be asked to tell the story in a sequence of pictures, where the pictures have been chosen to elicit the target words.

The previous paragraphs described work that follows immediately from the investigations carried out in this thesis. However, a limitation of the current approach is that while the findings may be *suggestive* of how articulatory errors influence spontaneous speech, they cannot establish causality. Therefore, there is a need for research that examines the direct effect of specific speech deficits on intelligibility. At the level of the single word, this is a relatively straightforward undertaking. For example, in Bunton and Weismer (2001) and Lansford and Liss (2014), perceptual assessment was used to categorise words into "intelligible" and "unintelligible" tokens. Different types of statistical analysis were then conducted to determine whether acoustic metrics were able to differentiate between these two categories. The design of explanatory studies in which the dependent variable is spontaneous-speech intelligibility is more challenging. For example, Zielinski (2006) investigated the determinants of spontaneous-speech intelligibility in a speaker of English as a second language. A transcript of the speaker's utterances was available such that the intended target was known. In addition to phonemic errors, she considered the effect of non-native syllable stress patterns on speech intelligibility. The speech sample was transcribed orthographically by three listeners and Zielinski compared these

transcriptions with the intended target. She then carried out a meticulous analysis procedure to identify non-native features that were truly *implicated in* (as opposed to just correlated with) breakdowns in intelligibility.

Zielinski's (2006) study demonstrates that even when the analysis is restricted to just two determinants of spontaneous-speech intelligibility, it can be challenging to tease out their relative importance. To give a simple example, an inaccurate vowel phoneme may in fact be a product of a non-standard syllable stress pattern. Thus the way in which disordered features interact and combine to mislead the listener may be quite complex. Given the large number of variables that may affect intelligibility in dysarthric speech, the applicability of Zielinski's (2006) approach to dysarthria research is likely to be limited.

Another option would be to conduct an experiment in which the effect of an articulatory error is studied in isolation. Klein and Flint (2006) asked a neurotypical speaker to create specific phonological errors (e.g., final consonant deletion or velar fronting) in a set of sentences. The direct effect of these errors on sentence intelligibility (word accuracy in an orthographic-transcription task) was then determined. The authors performed the experiment in two ways so as to determine the relative contributions of contrast vulnerability and functional load. Thus in the first experiment, the phonemic contrasts appeared at levels approximating those seen in conversational speech, while in the second, the incidence of each phonemic error was the same. Further studies of this kind could improve our understanding of the relative effects of different types of articulatory error on real-world intelligibility. However, the findings of manipulated-speech studies might have limited relevance to dysarthria for two reasons. Firstly, they do not take account of the aforementioned *interactions* between the various segmental and suprasegmental features. Secondly, they assume that speech errors are well-formed phonemic substitutions, an assumption that appears to be invalid, at least for most dysarthric misarticulations.

To bring this discussion to a close, it seems fitting to return to the central goal of impairment-based speech and language therapy for individuals with acquired dysarthria: to improve intelligibility in spontaneous speech. Accordingly, there is an urgent need for intervention studies that can directly assess whether correction of a particular articulatory deficit has the potential to improve functional intelligibility. In comparison with aphasia, there is a striking paucity of evidence for dysarthria interventions. According to a recent Cochrane review (Mitchell et al., 2017), there are no adequately-powered randomised controlled trials for speakers with dysarthria. Furthermore, in the studies that have been carried out, articulatory precision was generally addressed in an indirect manner, e.g., via speech-rate reduction or increased vocal effort – adjustments that are likely to affect all

subsystems of speech. Perhaps it is unrealistic to expect large clinical trials to focus on interventions of a highly specific nature, not least due to the difficulty in recruiting sufficient numbers of participants with the required deficit. However, given the current state of knowledge regarding articulatory treatment, even evidence from small-sample studies (e.g., Cahill et al., 2004; Robertson, 2001) would be of considerable value.

8.5. Summary and conclusions

This thesis described a series of investigations that aimed to (a) improve understanding of the methodological factors affecting the identification and categorisation of segmental speech errors in dysarthria by perceptual means, (b) examine the correlation between single-word reading intelligibility and intelligibility in spontaneous speech, and (c) obtain preliminary information about the phonemic errors of Antwerpian Dutch speakers with dysarthria. The main findings were as follows:

(1) The method of categorising phonemic errors according to contrasts in a single phonetic feature (Kent et al., 1989) shows considerable promise with regard to consonant confusions: at least 78% of the substitutions observed in any given speaker could be coded using 13 phonetic-contrast categories. For vowels, on the other hand, most of the individual confusions did not lend themselves to categorisation based on reasonably well-defined phonetic features (e.g., vowel height or backness).

(2) The following phonetic contrasts did not show evidence of being more vulnerable in speakers with dysarthria than in neurotypical speakers: the voicing of word-initial stops, nasal place confusions (both word-initial and word-final), and the directional vowel confusions /ɛ/ → /ɪ/ and /ɪ/ → /i/.

(3) Using the current version of the proposed single-word reading test, together with orthographic transcription to record listener responses, the cutoff word-accuracy scores for the diagnosis of dysarthria, based on data from neurotypical speakers, are 88.5% and 87.5%, for the 95% and 97.5% confidence levels respectively.

(4) Single-word reading accuracy was significantly higher for the four-alternative forced-choice mode than for orthographic transcription: the mean value of the absolute difference in the percentage of correct words (± 1 SD) was 13.1% ± 6.9%. Pearson's r between word-accuracy scores for the two response modes was high (r = 0.86, one-tailed $p < 0.01$), but speakers were not ranked in precisely the same order. For both vowel and consonant contrasts, the open and closed response modes showed differences in the top six error categories, which were defined on the basis of error-ranks calculated for each speaker and then summed over the cohort. For individual

speakers, the correlation between the ranked errors for the two response modes ranged from 0.34 - 0.72 (mean = 0.58) for consonant contrasts and 0.17 - 0.86 (mean = 0.47) for vowels.

(5) Reasonable correlation (r = 0.76, one-tailed p < 0.01) was obtained between a metric of spontaneous-speech intelligibility (Lagerberg et al., 2014) and initial-consonant accuracy in single-word reading. Informal perceptual assessment of the speakers' monologues revealed low numbers of clearly identifiable phonemic-substitution errors. There was insufficient evidence to draw definitive conclusions about the applicability of Lagerberg et al.'s (2014) method to speakers with dysarthria, but the conditions required for the method to be successful were discussed.

(6) The consonant contrast categories that were most vulnerable in the present cohort were initial-consonant devoicing, syllable-shape confusions (in particular, the perception of a cluster instead of a singleton at word-final position), and manner confusions, especially stop vs. fricative and /r/ vs. fricative (where the rhotic is an alveolar trill in the Antwerp accent). Place of articulation was not strongly affected. For vowels, there were durational errors in both directions (with vowel shortening being more common) and confusions between monophthongs and diphthongs. There was some preliminary evidence of possible compression of the vowel space in the front-back dimension, but this would need to be confirmed in future studies.

From a theoretical perspective, this thesis offers two main contributions. Firstly, it provides insights into speech production in individuals with mild to moderate dysarthria. Specifically, it adds to the body of evidence that dysarthric misarticulations are often mild distortions, and that when major distortions do occur, sufficient to be perceived as a substitution, they typically correspond to small interphonemic distances (unless heavily influenced by phonological and/or perceptual factors). For example, place of articulation errors tend to involve consonant pairs that have a fairly similar point of constriction, and vowel errors are generally confined to substitutions between phonemes that have been shown to occupy similar regions of the F1-F2 space for the accent in question. Furthermore, the majority of phonemic distortions in mild-moderate dysarthria involve just one phonetic feature. An exception to this statement concerns obstruent devoicing, which is a relatively common process in many languages of the world, including in neurotypical speakers. Therefore, it can often co-occur with a contrast in another phonetic feature.

Secondly, the findings of this thesis suggest a complex set of interactions between the individual's speech characteristics, the characteristics of the listener, and the methods used to elicit and assess speech. Of particular importance is the ability of the assessment to

differentiate between phonetic distortions and errors that cross a phoneme boundary. This thesis provided evidence to suggest that the extent to which a speaker's distortions might be perceived as substitutions, which could be regarded as 'false positive' outcomes, depends on both the nature of the speech stimulus and the method of recording the listener's responses. However, it is also possible that incorrect judgments will be made in the other direction, i.e., that a genuine substitution error will go undetected. This is more likely to occur in listener-response modes that involve a greater degree of constraint.

Appendix 1. Participant information sheet (English translation)

PARTICIPANT INFORMATION LEAFLET

Study of speech disorders due to brain injury

Full (technical) title: Relationship between segmental speech errors and intelligibility in speakers with acquired dysarthria

.

We would like to invite you to take part in a research study. Before you decide whether you would like to take part, it is important that you understand why the research is being done and what it would involve for you. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information.

What is the purpose of the study?

We plan to do a detailed analysis of the pronunciation problems of people with a brain injury. This will lead to better ways of assessing and treating speech difficulties in future patients. The duration of the study is 2 years and it is being undertaken as part of a PhD programme. It is a joint project between City University London (UK) and ZNA Middelheim (Belgium).

Why have I been invited?

The study will recruit 20 people with speech difficulties due to a brain injury. The cause of the brain injury must either be stroke or a disease of the cerebellum. These two groups were chosen because little is known about the speech difficulties of people with these conditions.

Do I have to take part?

It is up to you to decide whether or not to take part. If you do decide to take part, you will be asked to sign a consent form. After signing the form, you are still free to withdraw at any time.

What will happen if I take part?

You will participate in a video-recorded interview that takes about 40 minutes. The interview will take place in a quiet room of the Middelheim Hospital. It may be possible to conduct the interview over two sessions if this is your preference.

What do I have to do?

During the interview, the researcher will ask you to carry out some simple tasks, such as naming pictures and reading sentences. You will also have a brief conversation with the researcher about a subject that interests you. These speech tasks will allow us to identify your pronunciation errors.

What are the possible disadvantages and risks of taking part?

We do not expect participants to experience any negative effects from this study. The researcher who will be conducting the interview is a trained Speech and Language Therapist. She is aware of the fact that some people with speech difficulties can become tired or distressed during speaking activities. If she notices any negative effects, she will ask you whether you wish to stop the interview. You are also free to stop the interview yourself at any time and for any reason.

What are the possible benefits of taking part?

There are no direct benefits to the individual. However, information gained from this study will increase understanding about speech disorders due to brain injury, which will be helpful for future patients.

Will my taking part in the study be kept confidential?

Yes, we will follow ethical and legal practice. There are three items of personal data involved in this study, which will be handled as follows:

Firstly, the video recording of your interview is considered as personal data because you can be identified from it. The digital file of the recording will be kept in a secure, password-protected format and only the three researchers involved in this project will be given the access codes. The video data will only be used for the purposes of the present study.

Secondly, we will need to access your medical records so that we can make a note of any information relevant to this study (e.g. your type of brain injury). However, we will not write down any information that will allow you to be identified and our notes will be labelled only with a study number. Therefore, they will be anonymous. Your study number will be announced at the beginning of your interview so that we can link the interview with the medical notes.

Thirdly, your signed consent form contains personal data (your name and signature). This form will be kept in a locked filing cabinet in the office of the researcher at ZNA Middelheim.

What will happen when the research study stops?

After the study has been completed, your signed consent form will be securely destroyed. The digital file of your video recording will be stored on secure computers at City University London and ZNA Middelheim. They will be deleted after the time period required by

university guidelines. This may be a number of years. You have the right to obtain access to your recording using the contact details provided below.

What will happen to results of the research study?

We will publish the results as a PhD thesis and in scientific journals. These journals are accessible to the public (an access fee may apply). In any report or publication arising from this study, readers will not be able to identify any of the participants. We will prepare a summary sheet of the main findings. If you are still under the care of ZNA Middelheim, you will receive this sheet automatically. Otherwise, you can receive a copy by contacting us using the details below.

What will happen if I don't want to carry on with the study?

You are free to withdraw from the study at any time, without giving a reason. This will not affect your routine care or disadvantage you in any way. If you have already been interviewed, we will delete the file from our electronic storage system.

What if there is a problem?

If you have any problems, concerns or questions about this study, you should ask to speak to a member of the research team. If you remain unhappy and wish to complain formally, you can do this (a) through the Middelheim hospital complaints procedure, by telephoning [redacted] or by emailing [redacted], or (b) through the City University London complaints procedure, by telephoning [redacted] or by emailing [redacted].

City University London holds insurance policies which apply to this study. If you feel you have been harmed or injured by taking part in this study, you may be eligible to claim compensation. This does not affect your legal rights to seek compensation. If you are harmed due to someone's negligence, then you may have grounds for legal action.

Who has reviewed the study?

This study has been approved by the Research Ethics Committee of the School of Health Sciences, City University London and by the Medical Ethics Committee of ZNA Middelheim.

Further information and contact details

If you have any questions about any aspect of this study, please contact one of the investigators:

Naomi Miller: email - [redacted]

Johan Verhoeven: email - [redacted]

Peter Mariën: email - [redacted] or telephone [redacted].

Thank you for taking the time to read this information sheet.

Appendix 2. Participant consent form (English translation)

CONSENT FORM

Study of speech disorders due to brain injury

Full (technical) title: Relationship between segmental speech errors and intelligibility in speakers with acquired dysarthria

Participant Identification Number: ☐

Please initial box

1.	I agree to take part in the above research study, which will be carried out jointly by City University London and ZNA Middelheim. I have had the study explained to me, and I have read the Participant Information Leaflet, which I may keep for my records. I understand that the study will involve: • being interviewed by the researcher • allowing the interview to be videotaped • allowing the researchers to have access to my medical records	
2.	This information will be held and processed for the following purposes: • identifying speech difficulties • relating speech difficulties to medical diagnosis I understand that any information I provide is confidential, and that no information that could lead to the identification of any individual will be disclosed in any reports on the project, or to any other party. No identifiable personal data will be published. The identifiable data will not be shared with any other organisation.	
3.	I understand that my participation is voluntary, that I can choose not to participate in part or all of the project, and that I can withdraw at any stage of the project without being penalized or disadvantaged in any way.	
4.	I agree to City University London and ZNA Middelheim recording and processing this information about me. I understand that this information will be used only for the purposes set out in this statement and my consent is conditional on the researchers complying with its duties and	

	obligations under the Data Protection Act 1998 and Belgian Privacy Act 1992.	
5.	I agree to take part in the above study.	

_____	_____	_____
Name of Participant	Signature	Date

When completed, 1 copy for participant; 1 copy for researcher file; 1 copy for hospital notes.

Appendix 3. Target words and multiple-choice distractors

The following table lists the target words and multiple-choice foils employed in this study. When the English translation appears in bold typeface, this indicates that the word was also presented to the speakers as a picture. There was one word, /ʃaːl/, that was not tested in the multiple-choice study.

Word	*English*	*Foil 1*	*Foil 2*	*Foil 3*
boːt	**boat**	broːt	doːt	moːt
zɪŋ	(I) sing	zɪn	rɪŋ	vɪŋ
lɑs	(I) read (past tense)	ɣlɑs	lɔs	lɑst
rɔk	**skirt**	rɔx	roːk	ruk
vaːx	vague	vaːk	vraːx	vaːr
dɑm	dam	dɑn	tɑm	dɑrm
jɛi	you (emphatic)	hɛi	jɑ	zɛi
braːt	(I) roast	praːt	blaːt	baːt
toːn	(I) show	tɔn	tun	toːnt
wɔl	**wool**	wɔn	wɑl	wɔlf
køːs	choice	kʏs	kœys	knøːs
haːrt	hearth	hɑrt	aːrt	haːt
dut	(He) does	doːt	tut	but
veːr	**feather**	beːr	eːr	viːr
hʏt	hut	hut	heːt	hœyt
stiːr	**bull**	siːr	styːr	tiːr
doːf	deaf	dɔf	doːs	doːp
tɑk	**branch**	pɑk	dɑk	zɑk
nɛt	just, net	nit	nʏt	lɛt
yːr	hour	hyːr	uːr	pyːr
rœyt	window-pane, rhombus	raːt	zœyt	œyt
stɔp	(I) stop	sɔp	stɑp	stɔk
buːr	farmer	boːr	muːr	bux
hɔut	**wood** (the material)	ɔut	hœyt	ɣɔut
fɛl	intense	vɛl	sɛl	pɛl
zɑk	bag, pocket	zaːk	vɑk	zɑx
wɛn	(I) get used to	wɪn	rɛn	wɛin
ɣɔut	gold	ɣoːt	rɔut	hɔut

maːr	but	naːr	baːr	maːs
pɛin	pain	pɛn	peːn	fɛin
dɑk	**roof**	tɑk	bɑk	dɛk
spɪt	spit (as in roast)	spit	spɪts	pɪt
boːn	bean	bɔn	boːm	bun
ɣaːt	(He) goes	ɣɑt	raːt	haːt
heːn	to, forth	hɛin	eːn	heːt
pɪt	pip, seed (of a fruit)	pɛt	pit	pʏt
ʃaːl	scarf	-	-	-
bɔut	(He) builds	boːt	mɔut	baːt
oːx	**eye**	hoːx	oːr	loːx
vlis	membrane	vleːs	blis	flis
dɔf	dull	tɔf	doːf	bɔf
mɑxt	power, strength	nɑxt	mɑst	mɑx
wɪt	white	rɪt	wit	wɪs
zɔŋ	song	zɔn	zɑŋ	zɔŋk
pɛst	(He) bullies	bɛst	vɛst	tɛst
du	(I) do	tu	dul	dun
lɛist	list	rɛist	lʏst	leːst
ɣaːs	wire mesh	ɣɑs	haːs	raːs
jɑt	(He) pinches (as in steals)	ɣɑt	hɑt	jɑs
ʃɔk	shock	sɔk	jɔk	ɣɔk
vyːr	fire	byːr	vuːr	viːr
mɑt	mat, matt	nɑt	mɛt	bɑt
bɔt	bone	bɔts	bɑt	boːt
eːnt	duck	eːns	eːn	meːnt
plɑk	(I) stick	prɑk	plɛk	pɑk
kɪn	**chin**	kʏn	kɪnt	kin
wɛns	wish	wɛn	wɛnst	wɛnt
haːl	(I) collect, get	hɑl	aːl	haːr
meː	with	meːr	neː	mɛi
krɑm	clamp	kɑm	krɑmp	krɛm
roːt	**red**	rɔt	ɣroːt	ɣoːt
leːf	(I) am alive	neːf	bleːf	lɛif
nɔp	cleat	nɑp	noːp	knɔp
waːr	where, true	wɑr	waːx	raːr

hɑls	front of neck	ɑls	hɑlt	hɑlf
ɣrut	(I) greet	ɣro:t	ɣut	ɣlut
bɛnt	(You) are	bɪnt	bɑnt	pɛnt
lø:k	nice, fun	rø:k	lø:t	lœyk
trɑp	**stairs**	tɑp	trɑt	krɑp
zɪn	inclination, sense, sentence	zin	zɪŋ	vɪn
rø:s	giant	rœys	lø:s	hø:s
ko:rt	chord	ko:rts	o:rt	kɔrt
tɪl	(I) lift	tɛl	tɪn	stɪl
dra:x	(I) wear, carry	tra:x	da:x	dra:k
rʏk	(I) pull, jerk	rʏx	rø:k	drʏk
sɔp	sud (as in soap)	sɑp	sup	ɔp
lɛit	(He) leads	rɛit	lɛt	nɛit
ɛxt	real, really	hɛxt	vɛxt	ɛrt
ha:r	**hair**, her	a:r	ha:s	ha:rt
ɣɔk	guess	hɔk	rɔk	ɣɔt
by:r	neighbour	my:r	bø:r	dy:r
vɑŋ	(I) catch	bɑŋ	vɑn	zɑŋ
sxe:l	cross-eyed	sxɪl	sxe:r	sxe:lt
ɛi	egg	hɛi	ɛil	wɛi
ɣrɑf	grave	ɣrɑs	ɣrɔf	ɣra:f
pɛn	pen	bɛn	pɪn	pɛin
zɔut	salt	zœyt	vɔut	za:t
ho:r	(I) hear	o:r	ho:rt	ho:x
wɛist	(He) points	wɛst	wɛis	rɛist
bit	beetroot	pit	bɪt	be:t
ta:l	language	da:l	pa:l	ka:l
ne:r	down, low	de:r	ni:r	le:r
prɑt	proud	plɑt	prɛt	pra:t
ke:s	Kees (boy's name)	ka:s	kis	ke:t
fɛin	fine, good	sɛin	pɛin	fɛil
di:r	animal	bi:r	de:r	ni:r
krɔm	crooked	klɔm	krɑm	krɔp
hœyt	skin	œyt	rœyt	hʏt
rɛi	row, line, queue	zɛi	rɛim	vrɛi
stɛl	couple, (I) set (up)	stɪl	stɑl	sɛl

jɑs	jacket	ɣɑs	ɑs	jɑst
leːx	empty	lɪx	leːk	neːx
rɪt	reed	rɪt	reːt	ɣrɪt
ʃɔu	(I) haul	ʃoː	ʃɔut	zɔu
maːn	**moon**	baːn	mɑn	maːnt
vʏl	(I) fill	veːl	vul	ɣʏl
bœyt	booty	bɔut	bœys	mœyt
zɪŋk	zinc	vɪŋk	zɪn	hɪŋk
pɪl	**pill**	bɪl	pɛl	pʏl
krɑnt	**newspaper**	krɛnt	kɑnt	krɑns
miːr	**ant**	biːr	meːr	myːr
ɣeːl	**yellow**	heːl	ɣɪl	ɣeːn
dɑs	**badger**	tɑs	bɑs	daːs
bɛt	**bed**	bɪt	mɛt	bɑt
zaːx	**saw** (the tool)	zɑx	zaːxt	vaːx
rɛist	**rice**	rɛis	rɛst	rʏst
ɣrɑs	**grass**	ɣlɑs	ɣɑs	ɣrɑf

Appendix 4. Perceptual assessments of spontaneous speech

This appendix describes perceptual assessments of the monologues, as carried out by the author. The value in brackets after each speaker's ID represents their spontaneous-speech intelligibility. Note that the author has limited experience in the perceptual assessment of speech and is not a native speaker of Dutch. Therefore, the following observations are mainly restricted to substitution errors involving consonant phonemes. Occasional observations are made about other features (e.g., distortions, vowel errors, prosodic characteristics) when the author felt confident in her assessment. Instances of /h/-dropping are not reported. Finally, it is important to appreciate that transcripts of the monologues (i.e., the intended utterances) were not available. Therefore, it was only possible to make specific comments about articulatory errors within the *intelligible* portions of speech.

Speaker 1 (86.4%):

[sneβ̞] → [snet] or [snep]

No other substitutions were detected. However, in general, consonants were produced weakly and with slurring (i.e., movement from one articulation to another within the time of a single segment). For example, this occurred, in the phrase /də ˈrodə ˈbɑləcəs/. The speaker, who was later diagnosed with ALS, was originally suspected of having cerebellar disease based on the "slurred, drunken" quality of her speech, as assessed by a neurolinguist.

Speaker 2 (88.1%):

[plats] → [pats]

[ɣəˈlexə(n)] → [vəˈlexə]

No other obvious substitutions. General impression of weak, slurred sounds, e.g., in the word [vərˈsxɪləndə]. The speaker reported being tired after delivering the monologue.

Speaker 3 (88.1%):

[ˈɛixə(n)lək] → [ekələ]

[ˈbecə] → [ˈmyce]

[ɣəˈwest] → [ɣəˈweθ]

['darɔm] → ['warɔm]

[ɣə'krexə(n)] → [ɣə'krejə]

The speaker seemed to have vocal and prosodic deficits. His voice was weak and he had a vocal tremor and reduced loudness. He paused in inappropriate places.

Speaker 4 (86.5%):

Many instances of final cluster formation (thought to be a distortion rather than an intrusion error), e.g., [dun] → [dunt], [mɛn] → [mɛnt]. Otherwise, there were no *consistent* substitutions. The following isolated substitutions were observed:

[də] → [zə]

[hɛp] → [hɛm]

[dɪt] → [dɪs]

[dʏs] → [dʏx]

In addition, there seemed to be a large number of distortions that matched the substitutions observed in the free-response study. For example, /ɔ/ approached /ɑ/ in the word [stɔnt], /i:/ was distorted towards /y:/ in the word [vi:r], and most instances of /r/ were distorted such that they sounded somewhat like a fricative. In general, there was a "noisy, slushy" quality to some of her sounds, especially fricatives and /l/ (e.g., in the word ['lifstə]). It seemed that /t/ was sometimes partially fricated. However, none of these distortions would be classed as a clear substitution. (Note: in the case of word-final /r/, there is often an additional cue to the word-final phoneme. This is because /i, y, u/ are always lengthened before /r/, but not before other phonemes. Therefore it is less likely that a distortion of word-final /r/ would be heard as a substitution when preceded by one of these vowels, although word-final /r/ vs. fricative substitutions did arise in the assessment of single words).

Speaker 5 (91.2%):

['ɔpxədan] → ['ɔpxəɹan]

[ɣrotə] firstly pronounced as [hopə] and then partially corrected to [ɣropə]

[bə'drɛif] → [tə'rɛif]

[wɪlt] → [wult]

['ɪnrɪxtə(n)] → ['ɪndɪxtən]

No other substitutions were detected. However, the participant spoke very slowly and effortfully, which almost certainly reduced the number of errors. His main difficulty was polysyllabic words, which could become weakened and distorted. However, they rarely contained clearly definable substitution errors. He seemed to have a vocal deficit, most notably, variable pitch.

Speaker 6 (97.9%):

['heləmal] → ['hejəmal]

The remaining phonemic errors mainly consisted of deletions, e.g., [ɣə'drʏmt] → [ɣə'dʏmt], [sɪn(t)s'din] → [sɪn'sin], ['sxrɛivə(n)] → ['sxɛivə], [kwɑm] → [kɑn].

Otherwise, the overall impression was of weak, slurred sounds, as well as the telescoping of syllables. For example, in the phrase [ɪk **bɛn** ˈ**ɛi**xə(n)**lək**], which means "I am actually", only the syllables shown in boldface were clearly audible. However, some of these reductions could be considered dialect rather than a speech deficit.

Speaker 7 (96.8%):

No clear substitutions perceived. Possible dentalisation of /t/, e.g., in the word ['etə(n)].

Similarly to most of the other speakers, her deficit was characterised by generalised weakening, slurring and telescoping. When words were stressed or produced with effort, there was no discernible deficit. When a stream of multisyllabic words was uttered quickly and with lower effort, the individual words were slurred and telescoped, but the overall utterance was still highly intelligible, as can be seen from the high SSI score.

Speaker 8 (74.1%):

This speaker was identified as having cognitive-communication difficulties by the SLT team. After his first two utterances, he spoke very quickly and his monologue exhibited a very large amount of slurring, weakening, phoneme deletions, and telescoping. An example that lends itself to transcription was ['ɑltɛit ɔp] → [ɑl dup]. However, in general, an accurate transcription would be difficult. For example, in the word ['oxə(n)blɪk], which was completely unintelligible to the author but had been transcribed by some of the listeners, the velar fricative seemed to be deleted,

the bilabial plosive was produced with only very weak articulation, and the lateral approximant was deleted.

There also seemed to be a significant number of distortion errors. The speaker was unable to produce the alveolar trill and instead pronounced it as an alveolar approximant (English "r") or possibly even as [w]. This was a consistent error. Vowels, on the other hand, seemed to show *irregular* distortions, e.g., the vowel in [wet] sounded closer to [ɛi] and the vowel in [hœys] sounded closer to [o]. However, these did not seem to be full substitutions.

The speaker seemed to have prosodic deficits and some unusual vocal features – e.g., high pitch when emotional.

Speaker 9 (73.0%):

This participant was the speaker who exhibited excess and equal stress. He was the least intelligible speaker in the entire cohort (in spontaneous speech). This made it more difficult to identify specific substitution errors, as the intended target was often not known. Unlike all the other speakers, in his unintelligible portions of speech, the words were not slurred or uttered with reduced effort. Rather, the individual phonemes could be identified, but they did not correspond to any known Dutch words. In general, his speech seemed to contain lots of plosives, and the listeners had often transcribed words such as /də/ and /dɛn/ in contexts that did not make sense. The clear phonemic substitutions in the intelligible parts were:

/l/ in a cluster was consistently pronounced as an alveolar approximant (e.g., [blɛi] → [bɹɛi], [klɛin] → [kɹɛin]).

[mijn] → [bijn] (perceived on several occasions)

[vor] → [bor], [vɑn'avɔnt] → [bɑn'ovənt]

[zə] → [də]

['wedərɔm] → [ve'dɔm]

Deletions, such as [kɪnt] → [kɪn], [heft] → [het], [kɔmt] → [kɔm], [ɑls] → [ɑ], ['kɔntɛnt] → ['kɔntɛn]. The affix [-ə(n)], which denotes verb infinitives and past participles, was often omitted.

There were also lots of vowel distortions. It would be difficult to state definitively whether they were substitutions are not, but some were certainly heard that way according to the listeners' transcriptions. The most consistent examples were /ɛ/

distorted towards /ɑ/ (e.g., in ['bɛlə(n)], ['kɔntɛnt] and [zɛxt]) and /o/ distorted towards /u/ (e.g., in [ok] and ['sxonzon]).

Speaker 10 (92.2%):

['wɪnə(n)] → ['wɪdə]

[mer] showed partial denasalisation

[nit] → [lit]

Most of the errors yielded by this speaker were cases of generalised weakening and slurring, especially in multisyllabic words. His productions of /r/ seemed to be reduced to a tap or approximant, and the velar fricative was produced weakly, especially in medial position and unstressed syllables (e.g., in the words ['ɛixə(n)lək], [bə'xɔnə(n)] and [ɣə'stɔpt]).

References

Ackermann, H. & Hertrich, I. (1994). Speech rate and rhythm in cerebellar dysarthria: An acoustic analysis of syllable timing. *Folia Phoniatrica*, 46, 70-78.

Allison, K.M., Yunusova, Y. & Green, J.R. (2019). Shorter sentence length maximizes intelligibility and speech motor performance in persons with dysarthria due to amyotrophic lateral sclerosis. *Am J Speech Lang Pathol*, 28(1), 96-107.

Alves, M.O.C., Ode, C. & Strömbergsson, S. (2020). Dealing 0with the unknown – addressing challenges in evaluating unintelligible speech. *Clin Linguist Phon*, 34(1-2), 169-184.

Antolik, T.K. & Fougeron, C. (2013). Consonant distortions in dysarthria due to Parkinson's disease, amyotrophic lateral sclerosis and cerebellar ataxia. *Proceedings of Interspeech 2013*, 2152-2156.

Aronson, A.E. (1993). *Dysarthria: Differential Diagnosis (audio tape).* Rochester, MN: Mentor Seminars.

Beckman, M.E., Hirschberg, J. & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (Ed.) *Prosodic Typology – The Phonology of Intonation and Phrasing*, pp.9-54. Oxford: Oxford University Press.

Beijer, L.J., Rietveld, A.C.M, Ruiter, M.B. & Geurts, A.C.H. (2014). Preparing an E-learning-based Speech Therapy (EST) efficacy study: Identifying suitable outcome measures to detect within-subject changes of speech intelligibility in dysarthric speakers. *Clin Linguist Phon*, 28(12), 927-950.

Binger, C., Ragsdale, J. & Bustosa, A. (2016). Language sampling for preschoolers with severe speech impairments. *Am J Speech Lang Pathol*, 25(4), 493-507.

Black, J.W. (1969). *Relative Perceptual Similarity of Sixty Initial Consonants.* Columbus, OH: Ohio State University Research Foundation.

Blaney, B. & Hewlett, N. (2007). Dysarthria and Friedreich's ataxia: What can intelligibility assessment tell us? *Int J Lang Commun Disord*, 42(1), 19-37.

Boersma, P. & Weenink, D. (2018). *Praat: doing phonetics by computer* [Computer program]. Version 6.0.43, retrieved September 2018 from http://www.praat.org/.

Boothroyd, A. (2002). Context effects in spoken language perception. *Proc. Congreso Internacional de Foniatría, Audiología, Logopedia y Psicología del Lenguaje*. Universidad Pontificia de Salamanca.

Bosman, A.J. & Smoorenburg, G.F. (1995). Intelligibility of Dutch CVC syllables and sentences for listeners with normal hearing and with three types of hearing impairment. *Audiology*, 34(5), 260-284.

Brown, A. (1988). Functional load and the teaching of pronunciation. *Tesol Quarterly*, 22(4), 593-606.

Bunton, K., Leddy, M. & Miller, J. (2007). Phonetic intelligibility testing in adults with Down syndrome. *Downs Syndr Res Pract*, 12(1), 1-4.

Bunton, K. & Weismer, G. (2001). The relationship between perception and acoustics for a high-low vowel contrast produced by speakers with dysarthria. *J Speech Lang Hear Res*, 44(6), 1215-1228.

Bunton, K. & Weismer, G. (2002). Segmental level analysis of laryngeal function in persons with motor speech disorders. *Folia Phoniatr Logop*, 54(5), 223-239.

Cahill, L.M., Turner, A.B., Stabler, P.A., et al. (2004). An evaluation of continuous positive airway pressure (CPAP) therapy in the treatment of hypernasality following traumatic brain injury: A report of 3 cases. *J Head Trauma Rehabil*, 19(3), 241-253.

Campoy-Cubillo, M.C. (2016). Dysarthria and teaching speaking skills in English as a foreign language: A case study. *Miscelanea: A Journal of English and American Studies*, 53, 17-45.

CGN: Corpus Gesproken Nederlands (2018). Last accessed 30th March 2021 from: http://lands.let.ru.nl/cgn/.

Chakraborty, N. (2007). *A Linguistic Study of Dysarthric Bengali Speech*. PhD Thesis: University of Calcutta (Chapter 5). Last accessed 30th March 2021 from: https://shodhganga.inflibnet.ac.in/bitstream/10603/155893/11/11_chapter%205.pdf.

Clopper, C.G., Pisoni, D.B. & de Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *J Acoust Soc Am*, 118(3 Pt 1), 1661-1676.

Cohen, J.E. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Collins, B. & Mees, I.M. (2003). *The Phonetics of English and Dutch* (5th ed.). Leiden: Brill.

Coppens-Hofman, M.C., Terband, H., Snik, A.F.M., et al. (2016). Speech characteristics and intelligibility in adults with mild and moderate intellectual disabilities. *Folia Phoniatr Logop*, 68, 175-182.

Dagenais, P., Garcia, J. & Watts, C. (1998). Acceptability and intelligibility of mildly dysarthric speech by different listeners. In M.P. Cannito, K.M. Yorkston, D.R. Beukelman & P.H. Brookes (Eds.), *Neuromotor Speech Disorders: Nature, Assessment and Management* (pp. 229-239). Baltimore, MD: Paul H Brookes.

Dagenais, P.A., Watts, C.R.T., Turnage, L.M. & Kennedy, S. (1999). Intelligibility and acceptability of moderately dysarthric speech by three types of listeners. *Journal of Medical Speech-Language Pathology*, 7, 91-97.

Darley, F.L., Aronson, A. & Brown, J. (1969a). Differential diagnostic patterns of dysarthria. *J Speech Lang Hear Res*, 12, 246-269.

Darley, F.L., Aronson, A. & Brown, J. (1969b). Clusters of deviant speech dimensions in the dysarthrias. *J Speech Lang Hear Res*, 12, 462-496.

De Bodt, M., Guns, C., van Nuffelen, G., et al. (2006). *NSVO: Nederlandstalig SpraakVerstaanbaarheidsOnderzoek*. Vlaamse Vereniging voor Logopedisten (VVL).

De Bodt, M.S., Hernández-Díaz Huici, M.E. & van de Heyning, P.H. (2002). Intelligibility as a linear combination of dimensions in dysarthric speech. *J Commun Disord*, 35(3), 283-292.

De Louw, R. (2016). Is Dutch a pluricentric language with two centres of standardization? An overview of the differences between Netherlandic and Belgian Dutch from a Flemish perspective. *Werkwinkel*, 11(1), 113-135.

Duez, D. (2014). Some segmental and prosodic aspects of motor speech disorders in French. In N. Miller & A. Lowit (Eds.), *Motor Speech Disorders: A Cross-Language Perspective*, Ch. 12, pp.168-194. Bristol: Multilingual Matters Ltd.

Duffy, J.R. (2005). *Motor Speech Disorders: Substrates, Differential Diagnosis and Management* (2nd ed.). St Louis, Missouri: Elsevier Mosby.

Dutch 101. (2014). *1,000 Most Common Dutch Words*. Last accessed 30th March 2021 from: http://www.101languages.net/dutch/most-common-dutch-words/.

Enderby, P. & Palmer, R. (2008). *The Frenchay Dysarthria Assessment*. San Diego, CA: College Hill Press.

Eshghi, M., Richburg, B., Yunusova, Y. & Green, J.R. (2019). Instrumental evaluation of velopharyngeal dysfunction in amyotrophic lateral sclerosis. *Proceedings of International Congress of Phonetic Sciences ICPhS 2019*, 4-10 August 2019, Melbourne, Australia.

Fletcher, A.R., McAuliffe, M.J., Lansford, K.L. & Liss, J.M. (2017). Assessing vowel centralization in dysarthria: A comparison of methods. *J Speech Lang Hear Res*, 60, 341-354.

Flipsen, P. Jr., Hammer, J.B. & Yost, K.M. (2005). Measuring severity of involvement in speech delay: Segmental and whole-word measures. *Am J Speech Lang Pathol*, 14(4), 298-312.

Folker, J.E., Murdoch, B.E., Cahill, L.M., et al. (2010). Differentiating impairment levels in temporal versus spatial aspects of linguopalatal contacts in Friedreich's ataxia. *Motor Control*, 14, 490-508.

Gentil M. (1992). Phonetic intelligibility testing in dysarthria for the use of French language clinicians. *Clin Linguist Phon*, 6(3), 179-189.

Green, D.M. & Swets, J.A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

Haley, K.L., Bays, G.L. & Ohde, R.N. (2001). Phonetic properties of aphasic-apraxic speech: A modified narrow transcription analysis. *Aphasiology*, 15(12), 1125-1142.

Haley, K.L., Ohde, R.N. & Wertz, R.T. (2000). Single word intelligibility in aphasia and apraxia of speech: A phonetic error analysis. *Aphasiology*, 14(2), 179-201.

Haley, K.L., Smith, M. & Wambaugh, J.L. (2019). Sound distortion errors in aphasia with apraxia of speech. *Am J Speech Lang Pathol*, 28, 121-135.

Howard, S.J. & Heselwood, B.C. (2011). Instrumental and perceptual phonetic analysis: The case for two-tier transcriptions. *Clin Linguist Phon*, 25(11-12), 940-948.

Howard, S.J. & Heselwood, B.C. (2013). The contribution of phonetics to the study of vowel development and disorders. In M.J. Ball & F.E. Gibbon (Eds.), *Handbook of Vowels and Vowel Disorders*, pp.61-112. New York: Psychology Press.

Howson, P., Kochetov, A. & van Lieshout, P. (2015). Examination of the grooving patterns of the Czech trill-fricative. *J Phonetics*, 49, 117-129.

Hustad, K.C. (2007). Effects of speech stimuli and dysarthria severity on intelligibility scores and listener confidence ratings for speakers with cerebral palsy. *Folia Phoniatr Logop*, 59(6), 306-317.

Hustad, K.C., Dardis, C.M. & McCourt, K.A. (2007). Effects of visual information on intelligibility of open and closed class words in predictable sentences produced by speakers with dysarthria. *Clin Linguist Phon*, 21(5), 353-367.

IPA: International Phonetic Association (1999). *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press.

Johns, D.F. & Darley, F.L. (1970). Phonemic variability in apraxia of speech. *J Speech Lang Hear Res*, 13, 556-583.

Jongstra, W. (2003). *Variation in Reduction Strategies of Dutch Word-initial Consonant Clusters*. PhD Thesis: University of Toronto.

Jonkers, R., Terband, H. & Maassen, B. (2014). Diagnosis and therapy in adult acquired dysarthria and apraxia of speech in Dutch. In N. Miller & A. Lowit (Eds.), *Motor Speech Disorders: A Cross-Language Perspective*, Ch. 11, pp. 156-167. Bristol: Multilingual Matters Ltd.

Kawahara, S. & Garvey, K. (2014). Nasal place assimilation and the perceptibility of place contrasts. *Open Linguistics*, 1, 17-36.

Keintz, C.K., Bunton, K. & Hoyt, J.D. (2007). Influence of visual information on the intelligibility of dysarthric speech. *Am J Speech Lang Pathol*, 16(3), 222-234.

Kemper, S. & Sumner, A. (2001). The structure of verbal abilities in young and older adults. *Psychology and Aging*, 16(2), 312-322.

Kent, R.D. (1992). The biology of phonological development. In C. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological Development: Models, Research, Implications*, pp. 65-90. Timonium, MD: York Press.

Kent, R.D., Kent, J.F., Duffy, J.R., et al. (2000a). Ataxic dysarthria. *J Speech Lang Hear Res*, 43, 1275-1289.

Kent, J.F., Kent, R.D., Rosenbek, J.C., et al. (1992). Quantitative description of the dysarthria in women with amyotrophic lateral sclerosis. *J Speech Hear Res*, 35, 723-733.

Kent, R.D., Kent, J.F., Weismer, G. & Duffy, J.R. (2000b). What dysarthrias can tell us about the neural control of speech. *J Phonetics*, 28, 273-302.

Kent, R.D., Kent, J.F., Weismer, G., et al. (1990). Impairment of speech intelligibility in men with amyotrophic lateral sclerosis. *J Speech Hear Disord*, 55(4), 721-728.

Kent, R.D. & Rosenbek, J.C. (1983). Acoustic patterns of apraxia of speech. *J Speech Hear Res*, 26(2), 231-249.

Kent, R.D., Weismer, G., Kent, J.F. & Rosenbek, J.C. (1989). Toward phonetic intelligibility testing in dysarthria. *J Speech Hear Disord,* 54(4), 482-499.

Kent, R.D., Weismer, G., Kent, J.F., et al. (1999). Acoustic studies of dysarthric speech: methods, progress, and potential. *J Commun Disord*, 32, 141-186.

Kim, H., Kent, R.D. & Weismer, G. (2011). An acoustic study of the relationships among neurologic disease, dysarthria type, and severity of dysarthria. *J Speech Lang Hear Res*, 54(2), 417-429.

Kim, H., Martin, K., Hasegawa-Johnson, M. & Perlman, A. (2010). Frequency of consonant articulation errors in dysarthric speech. *Clin Linguist Phon*, 24(10), 759-770.

Klein, E.S. & Flint, C.B. (2006). Measurement of intelligibility in disordered speech. *Lang Speech Hear Serv Sch*, 37(3), 191-199.

Knuijt, S., Kalf J.G, van Engelen B.G., et al. (2017). The Radboud Dysarthria Assessment: Development and clinimetric evaluation. *Folia Phoniatr Logop*, 69, 143-153.

Kuruvilla-Dugdale, M., Custer, C., Heidrick, L., et al. (2018). A phonetic complexity-based approach for intelligibility and articulatory precision testing: A preliminary study on talkers with amyotrophic lateral sclerosis. *J Speech Lang Hear Res*, 61(9), 2205-2214.

Ladefoged, P. & Maddieson, I. (1996). *The Sounds of the World's Languages*. Oxford: Blackwell.

Lagerberg, T.B., Asberg, J., Hartelius, L. & Persson, C. (2014). Assessment of intelligibility using children's spontaneous speech: Methodological aspects. *Int J Lang Commun Disord*, 49(2), 228-239.

Lansford, K. & Liss, J. (2014). Vowel acoustics in dysarthria: Mapping to perception. *J Speech Lang Hear Res*, 57, 68-80.

Lee, J., Hustad, K.C. & Weismer, G. (2014). Predicting speech intelligibility with a multiple speech subsystems approach in children with cerebral palsy. *J Speech Lang Hear Res*, 57(5), 1666-1678.

Lillvik, M., Allemark, E., Karlström, P. & Hartelius, L. (1999). Intelligibility of dysarthric speech in words and sentences: Development of a computerized assessment procedure in Swedish. *Logop Phoniatr Vocol*, 24, 107-119.

Lindblom, B. (1990). On the communication process: Speaker listener interaction and the development of speech. *Augment Altern Commun*, 6, 220-230.

Liss, J.M., White, L.S., Mattys, S., et al. (2009). Quantifying speech rhythm abnormalities in the dysarthrias. *J Speech Lang Hear Res*, 52, 1334-1352.

Logemann, J.A. & Fisher, H.B. (1981). Vocal tract control in Parkinson's disease: Phonetic feature analysis of misarticulations. *J Speech Hear Disord*, 46, 348-352.

Luyckx, K., Kloots, H., Coussé, E. & Gillis, S. (2007). Klankfrequenties in het Nederlands. In D. Sandra (Ed.), *Tussen Taal, Spelling en Onderwijs. Essays bij het Emeritaat van Frans Daems*, pp.141-154. Gent: Academia Press.

Mackenzie, C. (2011). Dysarthria in stroke: a narrative review of its description and the outcome of intervention. *Int J Speech Lang Pathol*, 13(2), 125-136.

McHenry, M. (2011). An exploration of listener variability in intelligibility judgments. *Am J Speech-Lang Pathol*, 20, 119-123.

McHugh, M.L. (2012). Interrater reliability: The kappa statistic. *Biochem Med (Zagreb)*, 22(3), 276-282.

McLoughlin, V. (2009). *Applied Speech and Audio Processing*. Cambridge: Cambridge University Press.

Martens, H., van Nuffelen, G., van den Putte, L., et al. (2010). Meten van spraakverstaanbaarheid op zinsniveau bij volwassenen met een spraakstoornis: Introductie van het Nederlandstalig spraakverstaanbaarheidsonderzoek-zinsniveau (NSVO-Z). *Logopedie*, 2, 21-26.

Memrise (2014). *Dutch – The 1001 Most Common Words.* Dutch Lexicon Project, Dept. of Experimental Psychology at the University of Gent. Last accessed 30th March 2021 from: http://www.memrise.com/course/408/dutch-the-1001-most-common-words/.

Miller, N. (1995). Pronunciation errors in acquired speech disorders: The errors of our ways. *Eur J Disord Commun*, 30(3), 346-361.

Miller, N. (2013). Measuring up to speech intelligibility. *Int J Lang Commun Disord*, 48(6), 601-612.

Miller, J.L. & Baer, T. (1983). Some effects of speaking rate on the production of /b/ and /w/. *J Acoust Soc Am*, 73(5), 1751-1755.

Mitchell, C., Bowen, A., Tyson, S., et al. (2017). Interventions for dysarthria due to stroke and other adult-acquired, non-progressive brain injury. *Cochrane Database of Systematic Reviews 2017*, 1.

Murdoch, B.E. (2011). Physiological investigation of dysarthria: Recent advances. *Int J Speech-Lang Pathol*, 13(1), 28-35.

Narayan, C.R. (2008). The acoustic-perceptual salience of nasal place contrasts. *J Phonetics*, 36(1), 191-217.

Nordli, I.C. (1996). *Ataxic /r/ - Articulatory Description and Treatment Using Electropalatography (EPG): A Case Study*. Report retrieved from Munin Open Research Archive. Last accessed 30th March 2021 from: http://hdl.handle.net/10037/8934.

Odell, K., McNeil, M.R., Rosenbek, J.C. & Hunter, L. (1991). Perceptual characteristics of vowel and prosody production in apraxic, aphasic and dysarthric speakers. *J Speech Hear Res*, 34(1), 67-80.

Oldfield, R.C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropyschologia*, 9(1), 97-113.

Palmer, R. (2005). *An Evaluation of Speech and Language Therapy for Chronic Dysarthria: Comparison of Conventional and Computational Approaches*. Thesis: Institute of General Practice and Primary Care, University of Sheffield.

Palmer, R. & Enderby, P. (2007). Methods of speech therapy treatment for stable dysarthria: A review. *Int J Speech-Lang Pathol*, 9(2), 140-153.

Patel, R., Usher, N., Kember, H., et al. (2014). The influence of speaker and listener variables on intelligibility of dysarthric speech. *J Commun Disord*, 51, 13-18.

Platt, L.J., Andrews, G., Young, M. & Quinn, P.T. (1980a). Dysarthria of adult cerebral palsy: I. Intelligibility and articulatory impairment. *J Speech Hear Res*, 23(1), 28-40.

Platt, L.J., Andrews, G. & Howie, P.M. (1980b). Dysarthria of adult cerebral palsy: II. Phonemic analysis of articulation errors. *J Speech Hear Res*, 23(1), 41-55.

Pols, L.C.W. (1983). Three-model principal component analysis of confusion matrices, based on the identification of Dutch consonants, under various conditions of noise and reverberation. *Speech Communication*, 2, 275-293.

Poundstone, W. (2015). *Rock Breaks Scissors: A Practical Guide to Outguessing and Outwitting Almost Everybody*. New York: Little, Brown and Company.

Pye, C., Ingram, D. & List, H. (1987). A comparison of initial consonant acquisition in English and Quiché. In K. Nelson & A. van Kleek (Eds.), *Children's Language, Vol. 6*, pp.175-190. Hillsdale: Erlbaum.

RCSLT: Royal College of Speech and Language Therapists (2009). *RCSLT Resource Manual for Commissioning and Planning Services for SLCN*.

Read, J., Miller, N. & Kitsou, N. (2018). Is there an order of loss of sounds in speakers with Parkinson's disease? *Clin Linguist Phon*, 32(11), 997-1011.

Redford, M.A. & Diehl, R.L. (1999). The relative perceptual distinctiveness of initial and final consonants in CVC syllables. *J Acoust Soc Am*, 106, 1555-1565.

Riddell, J., McCauley, R., Mulligan, M. & Tandan, R. (1995). Intelligibility and phonetic contrast errors in highly intelligible speakers with amyotrophic lateral sclerosis. *J Speech Hear Res*, 38(2), 304-314.

Robertson, S. J. (1982). *Dysarthria Profile*. Bicester: Winslow.

Robertson, S.J. (2001). The efficacy of oro-facial and articulation exercises in dysarthria following stroke. *Int J Lang Commun Disord*, 36(S1), 292-297.

Rong, P.Y.Y., Yunusova, J., Wang, L., et al. (2016). Predicting speech intelligibility decline in amyotrophic lateral sclerosis based on the deterioration of individual speech subsystems. *PLoS One*, 11(5): e0154971.

Rudzicz, F. (2011). Acoustic transformations to improve the intelligibility of dysarthric speech. *Proc. 2nd Workshop on Speech and Lang. Processing for Assistive Technologies*, 30 July, Edinburgh, Scotland.

Schiavetti, N. (1992). Scaling procedures for the measurement of speech intelligibility. In R. Kent (Ed.), *Intelligibility in Speech Disorders*, pp. 11-34. Philadelphia, PA: John Benjamins.

Schmahmann, J.D. & Sherman, J.C. (1998). The cerebellar cognitive affective syndrome. *Brain*, 121, 561-579.

Schmidt, R.A. (1975). A schema theory of discrete motor skill learning. *Psychological Review*, 82, 225-260.

Sebregts, K. (2015). *The Sociophonetics and Phonology of Dutch r*. Dissertation: Utrecht University Repository. Last accessed 30th March 2021 from: https://dspace.library.uu.nl/handle/1874/306415.

Shriberg, L. & Kent, R. (1982). *Clinical Phonetics*. New York: Macmillan.

Shriberg, L.D. & Kwiatkowski, J. (1982). Phonological disorders III: A procedure for assessing severity of involvement. *J Speech Hear Disord*, 47, 256-270.

Shriberg, L., Kwiatkowski, J. & Hoffman, K. (1984). A procedure for phonetic transcription by consensus. *J Speech Hear Disord*, 51, 309-324.

Smith, C.H., Patel, S., Woolley, R.L., et al. (2019). Rating the intelligibility of dysarthic speech amongst people with Parkinson's Disease: A comparison of trained and untrained listeners. *Clin Linguist Phon*, 33(10-11), 1063-1070.

Stipancic, K.L., Tjaden, K. & Wilding, G. (2016). Comparison of intelligibility measures for adults with Parkinson's disease, adults with multiple sclerosis, and healthy controls. *J Speech Lang Hear Res*, 59(2), 230-238.

Stokes, S.F. & Surendran, D. (2005). Articulatory complexity, ambient frequency, and functional load as predictors of consonant development in children. *J Speech Lang Hear Res*, 48(3), 577-591.

Tjaden, K., Sussman, J.E. & Wilding, G.E. (2014). Impact of clear, loud, and slow speech on scaled intelligibility and speech severity in Parkinson's disease and multiple sclerosis. *J Speech Lang Hear Res*, 57(3), 779-792.

Tjaden, K. & Wilding, G. (2004). Rate and loudness manipulations in dysarthria: Acoustic and perceptual findings. *J Speech Lang Hear Res*, 47(4), 766-783.

Tjaden, K. & Wilding, G. (2011) Effects of speaking task on intelligibility in Parkinson's disease. *Clin Linguist Phon*, 25(2), 155-168.

Tomić, D. & Mildner, V. (2015). Development of /r/ in Croatian. *Proceedings of the 18th International Congress of Phonetic Sciences*. The Scottish Consortium for ICPhS 2015 (ur.). Glasgow, UK: The University of Glasgow, 2015.

Tomik, B. & Guiloff, R.J. (2010). Dysarthria in amyotrophic lateral sclerosis: A review. *Amyotroph Lateral Scler*, 11(1-2), 4-15.

Urban, P.P., Rolke, R., Wicht, S., et al. (2006). Left-hemispheric dominance for articulation: A prospective study on acute ischaemic dysarthria at different localizations. *Brain*, 129, 767-777.

Van Alphen, P.M. & Smits, R. (2004). Acoustical and perceptual analysis of the voicing distinction in Dutch initial plosives: The role of prevoicing. *J Phonetics*, 32, 455-491.

Van Dale Uitgevers (2009). *Middelgroot Woordenboek Nederlands-Engels* (1st ed.). Utrecht / Antwerpen: Van Dale Uitgevers.

Van Nuffelen, G., de Bodt, M., Guns, C., et al. (2008). Reliability and clinical relevance of segmental analysis based on intelligibility assessment. *Folia Phoniatr Logop*, 60, 264-268.

Van Nuffelen, G., de Bodt, M., Wuyts, F. & van de Heyning, P. (2009a). The effect of rate control on speech rate and intelligibility of dysarthric speech. *Folia Phoniatr Logop*, 61, 69-75.

Van Nuffelen, G., Middag, C., de Bodt, M. & Martens, J-P. (2009b). Speech technology-based assessment of phoneme intelligibility in dysarthria. *Int J Lang Comm Dis*, 44(5), 716-730.

Van Severen, L., Gillis, J.J.M, Molemans, I., et al. (2013). The relation between order of acquisition, segmental frequency and function: The case of word initial consonants in Dutch. *Journal of Child Language*, 40(4), 703-740.

Vandana, V.P. & Manjula, R. (2015). Speech intelligibility in ataxic dysarthria due to lesions in different cerebellar loci. *Language in India*, 15(5), 381-390.

Verhoeven, J. (2005). Belgian Standard Dutch. *J Int Phon Assoc*, 35(2), 243-247.

Verhoeven, J. & Hageman, G. (2007). De verstemlozing van fricatieven in Vlaanderen. *Nederlandse Taalkunde*, 12, 139-152.

Verhoeven, J. & van Bael, C. (2002). Akoestische kenmerken van de Nederlandse klinkers in drie Vlaamse regio's. *Taal en Tongval*, 54, 1-23.

Verkhodanova, V. & Coler, M. (2018). Prosodic and segmental correlates of spontaneous Dutch speech in patients with Parkinson's disease: A pilot study. Paper presented at *9th Speech Prosody Conference*, Poznan, Poland, 13th -16th June 2018. Last accessed 30th March 2021 from: https://www.isca-speech.org/archive/SpeechProsody_2018/pdfs/98.pdf.

Vigouroux, J. & Miller, N. (2007). Intelligibility testing: Issues in closed versus open format scoring. *Newcastle and Durham Working Papers in Linguistics*, 12, 83-95.

Walshe, M. & Miller, N. (2011). Living with acquired dysarthria: The speaker's perspective. *Disabil Rehabil*, 33, 195-203.

Warner, N., Smits, R., McQueen, J.M. & Cutler, A. (2005). Phonological and statistical effects on timing of speech perception: Insights from a database of Dutch diphone perception. *Speech Commun*, 46, 53-72.

Weismer, G., Jeng, J-Y., Laures, J.S., et al. (2001). Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders. *Folia Phoniatr Logop*, 53, 1-18.

Weismer, G. & Laures, J.S. (2002). Direct magnitude estimates of speech intelligibility in dysarthria: Effects of a chosen standard. *J Speech Lang Hear Res*, 45(3), 421-433.

Weismer, G. & Martin, R.E. (1992). Acoustic and perceptual approaches to the study of intelligibility. In R.D. Kent (Ed.), *Intelligibility in Speech Disorders: Theory, Measurement, and Management*, pp. 67-118. Philadelphia: John Benjamins.

Westbury, J.R. & Keating, P.A. (1986). On the naturalness of stop consonant voicing. *J Linguistics*, 22(1), 145-166.

Whitehill, T.L. & Ciocca, V. (2000a). Speech errors in Cantonese speaking adults with cerebral palsy. *Clin Linguist Phon*, 14(2), 111-130.

Whitehill, T.L. & Ciocca, V. (2000b). Perceptual-phonetic predictors of single-word intelligibility: A study of Cantonese dysarthria. *J Speech Lang Hear Res*, 43, 1451-1465.

Whitehill, T.L., Ciocca, V. & Yiu, E.M.L. (2004). Perceptual and acoustic predictors of intelligibility and acceptability in Cantonese speakers with dysarthria. *Journal of Medical Speech-Language Pathology*, 12(4), 229-233.

Wilson, E.M., Abbeduto, L., Camarata, S.M. & Shriberg, L.D. (2019). Speech and motor speech disorders and intelligibility in adolescents with Down syndrome. *Clin Linguist Phon*, 33(8), 790-814.

Xue, W., Cucchiarini, C., van Hout, R.W.N.M. & Strik, H. (2019). Acoustic correlates of speech intelligibility. The usability of the eGeMAPS feature set for atypical speech. In: *Proceedings of SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education*, pp. 48-52.

Yorkston, K.M. & Beukelman, D.R. (1978). A comparison of techniques for measuring intelligibility of dysarthric speech. *J Commun Disord*, 11, 499-512.

Yorkston, K.M. & Beukelman, D.R. (1980). A clinician-judged technique for quantifying dysarthric speech based on single-word intelligibility. *J Commun Disord*, 13, 15-31.

Yorkston, K.M. & Beukelman, D.R. (1981). *Assessment of Intelligibility of Dysarthric Speech*. Tigard, OR: CC Publ.

Yorkston, K.M., Beukelman, D.R. & Bell, K.R. (1987). *Clinical Management of Dysarthric Speakers*. London: Taylor & Francis.

Yorkston, K.M., Beukelman, D.R., Hakel, M. & Dorsey, M. (2007). *Speech Intelligibility Test for Windows*. Lincoln, NE: Institute for Rehabilitation Science and Engineering at Madonna Rehabilitation Hospital.

Yunusova, Y., Weismer, G., Kent, R.D. & Rusche, N.M. (2005). Breath-group intelligibility in dysarthria: Characteristics and underlying correlates. *J Speech Lang Hear Res*, 48, 1294-1310.

Ziegler, W. (2016). The phonetic cerebellum: Cerebellar involvement in speech sound production. In P. Mariën & M. Manto (Eds.), *The Linguistic Cerebellum*, pp.1-32. London, UK: Academic Press.

Zielinski, B. (2006). The intelligibility cocktail: An interaction between speaker and listener ingredients. *Prospect: An Australian Journal of TESOL*, 21(1), 22-45.

Zraick, R.I. & Liss, J.M. (2000). A comparison of equal-appearing interval scaling and direct magnitude estimation of nasal voice quality. *J Speech Lang Hear Res*, 43(4), 979-988.

Zuidema, W. (2009). *A Syllable Frequency List for Dutch*. Last accessed 30th March 2021 from: http://www.illc.uva.nl/Research/Reports/PP-2009-50.text.pdf.