Jérémy Genette

Universiteit Antwerpen

CLiPS Computational Linguistics & Psycholinguistics Research Center

Linguistics Departement

Stadscampus - L 305

Prinsstraat 13 - 2000 Antwerpen

jeremy.genette@uantwerpen.be

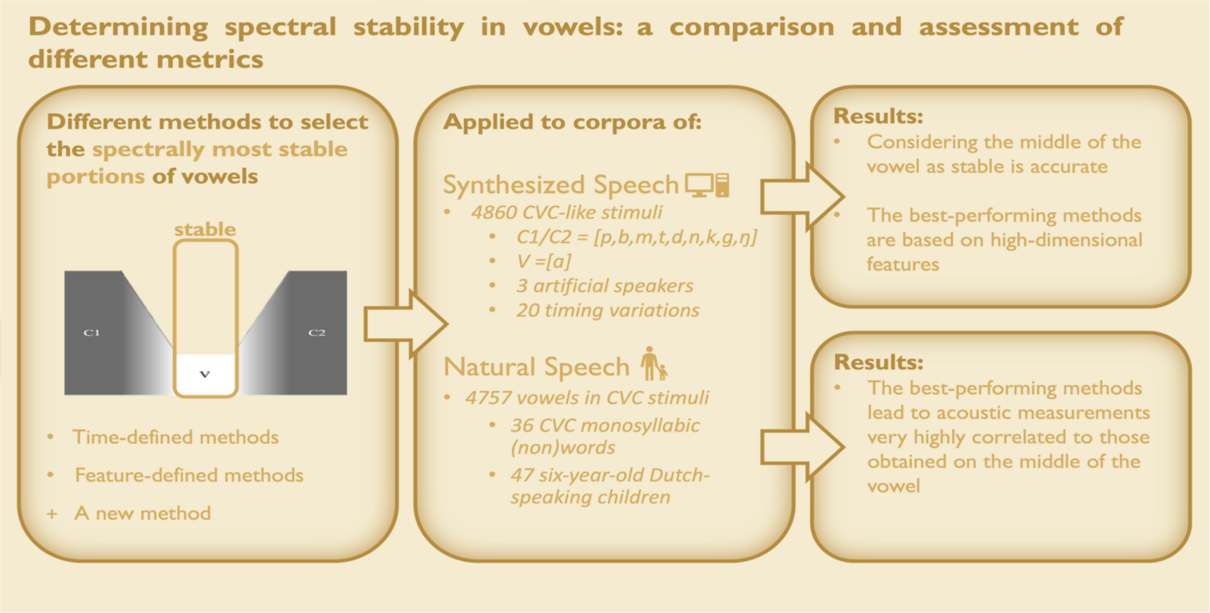
Determining spectral stability in vowels: a comparison and assessment of different metrics

Jérémy Genette[a], Jose Manuel Rivera Espejo[b], Steven Gillis[a], Jo Verhoeven[a,c]

[a] *Universiteit Antwerpen, Department of Linguistics, CLIPS Computational Linguistics and Psycholinguistics, Antwerp, Belgium*
[b] *Department of Training and Education Sciences, Research Group Edubron, Universiteit Antwerpen, Antwerp, Belgium*
[c] *Department of Language and Communication Science, City, University of London, London, United Kingdom*

Abstract

This study investigated the performance of several metrics used to evaluate spectral stability in vowels. Four metrics suggested in the literature and a newly developed one were tested and compared to the traditional method of associating the spectrally stable portion with the middle of the vowel. First, synthetic stimuli whose spectrally stable portion had been defined in advance were used to evaluate the potential of the different metrics to capture spectral stability. Second, the output of the different metrics on the acoustic measurements obtained in the vowel portions identified as spectrally stable was compared on both synthesized and natural speech. It is clear that higher-dimensional features are needed to capture spectral stability and that the best-performing metrics yield acoustic measurements that are similar to those obtained in the middle of the vowel. This study empirically validates long-standing intuitions about the validity of selecting the middle section of vowels as the preferred method to identify the spectrally stable region in vowels.



Determining spectral stability in vowels: a comparison and assessment of different metrics

Different methods to select the spectrally most stable portions of vowels

- Time-defined methods
- Feature-defined methods

+ A new method

Applied to corpora of:

Synthesized Speech

- *4860 CVC-like stimuli*
  - *C1/C2 = [p,b,m,t,d,n,k,g,ŋ]*
  - *V =[a]*
  - *3 artificial speakers*
  - *20 timing variations*

Natural Speech

- *4757 vowels in CVC stimuli*
  - *36 CVC monosyllabic (non)words*
  - *47 six-year-old Dutch-speaking children*

Results:

- Considering the middle of the vowel as stable is accurate
- The best-performing methods are based on high-dimensional features

Results:

- The best-performing methods lead to acoustic measurements very highly correlated to those obtained on the middle of the vowel

Keywords: Spectral analysis, Methodology, Vowel target, Feature-selection, Speech synthesis

1. Introduction

Speech sounds are categorized by speakers into discrete units which support communication and many investigations of speech aim to measure the acoustic characteristics of those units. For practical reasons, the dynamics of vowels have often been reduced to a single measurement per vowel which is considered representative for the whole vowel even if the discrete linguistic units are mapped onto a continuous acoustic signal. However, that task is anything but trivial. In fact, selecting a specific measurement point inevitably influences the characterization of the vowel to some extent by phenomena such as co- articulation (e.g. Kühnert and Nolan, 1999; Farnetani and Recasens, 2010; Embarki and Dodane, 2011).

Selecting a measurement point which is minimally influenced by the phonetic context is thus not a trivial task and it involves identifying the spectrally most stable portion,[1] which in earlier studies was done by visual inspection of spectrograms. However, with phonetics resorting to ever larger databases, there is a need for automating this process. To this end, several metrics have been suggested in the literature to evaluate spectral stability of vowels automatically. These metrics can be broadly classified into two types (Evanini, 2009: 5), i.e. a distinction can be made between time-defined and feature-defined methods. The former are based on the analysis of the time dimension, while the latter rely on the examination of acoustic features. Furthermore, feature-defined methods can be vowel-independent or vowel-dependent, i.e. the acoustic features are specific to the type of vowel involved.

These approaches have advantages and disadvantages and it is of interest to assess the different techniques available to identify the spectrally stable portion of vowels. In the remainder of this paper, this task will be referred to as Identification of Maximal Spectral Stability (IMSS). To the best of our knowledge, only Evanini (2009) has compared different methods for IMSS. Evanini (2009: 66) suggests that feature-defined methods such as those of Lennig (1978) and Labov et al. (2006) yield inferior results than time-defined methods. When applying time-defined methods, the spectrally stable portion of vowels is situated around the first quarter or third of the total vowel duration. However,

Evanini's (2009) evaluation has two major drawbacks. Firstly, the comparison only included a limited number of metrics. Secondly, it is based on the degree to which the formant measurements correlate with those obtained after manual selection. The problem with using human-made selections as a reference is that the evaluation procedure is based on the reliability of the reference. However, the reliability of a manual selection is difficult to assess because the criteria used by phoneticians are often vague and the consistency of their application is unknown (Van Bergem, 1988: 62).

Therefore, a more reliable benchmark to compare the different techniques is required. This is precisely what this study intends to achieve.

[1] Please note that the phonetic context can significantly influence even the most spectrally stable portion of the vowel. However, one should search for the segment that is relatively less affected by the preceding and following phones in comparison to other parts of the vowel, such as the transitions between the vowel and its adjacent phones. In other words, it might be impossible to find a portion of the vowel which is not under the influence of its phonetic context, but the aim is to minimize such contextual effects by excluding transitional phases and selecting the segment with the highest spectral stability.

2. Background

The most intuitive approach for identifying the spectrally stable portion of an acoustic signal is to rely on visual inspection of sound spectrograms. However, this technique has two major drawbacks. Firstly, the criteria are unclear, and the consistency of their application cannot be assured. Secondly, the increasing size of databases available to phoneticians makes visual inspection nearly unfeasible. For these reasons, it has been attempted to automate this process.

In the past, several procedures have been proposed for automatic IMSS. Perhaps the most widely used technique is one that selects the middle of the vowel. That technique relies on a commonly accepted model of the organization of articulatory transitions according to which a vowel phoneme can be considered as consisting of three parts: the vowel onset, a spectrally stable portion and the vowel offset (Lindblom and Studdert-Kennedy, 1967: 831). Typically, selecting the middle third of the vowel is considered a good approach. Because this time-defined technique is widely used in phonetic sciences, it serves as a reference for comparison throughout this paper and is operationalised here as the portion of the vowel situated at its temporal center and whose duration is equal to one-third of the total duration of the vowel.

The main advantage of this technique is its ease of implementation. However, a study by Weismer and Berry (2003) suggests that there might be significant individual differences in formant dynamics which suggests that *a priori* selecting the center of vowels may fail to identify the stable portion of vowels.

In order to deal with this, it has been attempted to develop metrics which evaluate spectral stability to identify the most stable portion in vowels. Several techniques have been proposed for this purpose, but they all rely on the same basic principle. If the acoustic signal is composed of successive frames, a given acoustic feature can be measured in each of them. Then, the extent of dissimilarity between the extracted features in the different frames is compared and the most dissimilar sequences of frames are excluded. The most important advantage of such feature-defined algorithms is that they do not rely on *a priori* assumptions about the location of the stable portion of vowels.

For instance, Lennig (1978: 52–53) suggested a metric which selects the portion of the signal where “the spectrum at which the first two formants [are] changing the least quickly”. In other words, it computes an instability score. Based on a perceptual study by Miller (1989), Hillenbrand et al. (1995: 3100) suggested a similar technique to identify the spectrally most stable portion of a vowel by calculating the slope in log F2 - log F1. Van Bergem (1988) developed another metric to evaluate frame-per-frame dissimilarity consisting of the pooled within-variance of the log-transformed formant values in each frame. The main disadvantage of these techniques is that they use the stability of the formants as a proxy for the overall spectral stability, leading to a low-dimensional representation of spectral stability. A second drawback of formant-based techniques is that the identification of spectral stability relies exclusively on the accuracy of the formant tracking. However, it is well known that automatic formant tracking can often be biased (Van der Harst, 2011).

Van Bergem (1993: 6) developed another metric which is not based on formant values but on mel-like scale cepstral coefficients (also called MFCC), which are regularly used in speech processing. Leaving aside technical details, the features used by speech processing applications have a number of attractive characteristics. They usually do not use a single acoustic feature but a combination of them, either at the level of the feature vector itself or at the level of acoustic probabilities (cfr. Nadeu et al., 2001: 516). The second advantage of speech-processing techniques is that they use features whose window of analysis is not as limited in frequency range as formant-based techniques.

However, the main disadvantage of speech processing approaches to the identification of spectral stability is that they are designed to search for changes between phones, i.e. to chunk the signal into contrastive units only. In other words, they need to capture dissimilarities in the acoustic signal which are relevant to the human ear. Therefore, the spectral feature(s) need to be sensitive to ensure that the algorithm can detect a boundary between two phones but stable within a single segment for the algorithm not to detect boundaries within a single segment.

Besides those techniques, there are also vowel-dependent techniques which use different features for each vowel category (e.g. Labov et al., 2006; Fletcher et al., 2015; Derdemezis et al., 2016, Fletcher et al., 2017; Eichhorn et al., 2018). Therefore, they require pre-processing in the form of transcriptions and cannot be applied to speech sounds not suitable for phoneme-based classification such as babbling. Vowel-dependent techniques are beyond the scope of the present paper.

Overall, the present review shows that the different IMSS procedures have advantages and disadvantages (a more formal characterisation of the different techniques can be found in the appendix A). In order to further improve the IMSS, the method needs to be flexible, cover a wide range of frequencies and does not rely on a potentially inaccurate tracking algorithm. It was consequently decided to try and develop a new metric to identify the spectrally most stable portion of vowel sounds.

3. A new metric

This section introduces the Spectral Stability Score (SSS) which is a new metric to identify spectral stability: its basic principles are illustrated in Fig. 1. This metric operates on the idea that spectral stability is more accurately represented when large(r) frequency ranges are analysed, when few(er) transformations are applied to the acoustic signal and when no complex tracking algorithm is used. To meet those requirements, the present SSS uses the Long-Term Average Spectrum (henceforth, LTAS). The LTAS has long been used as a proper “long- term” spectral feature because it is considered highly invariant with respect to segmental influences when speech samples exceed 20–40 s (for a review, see Mennen et al., 2010).

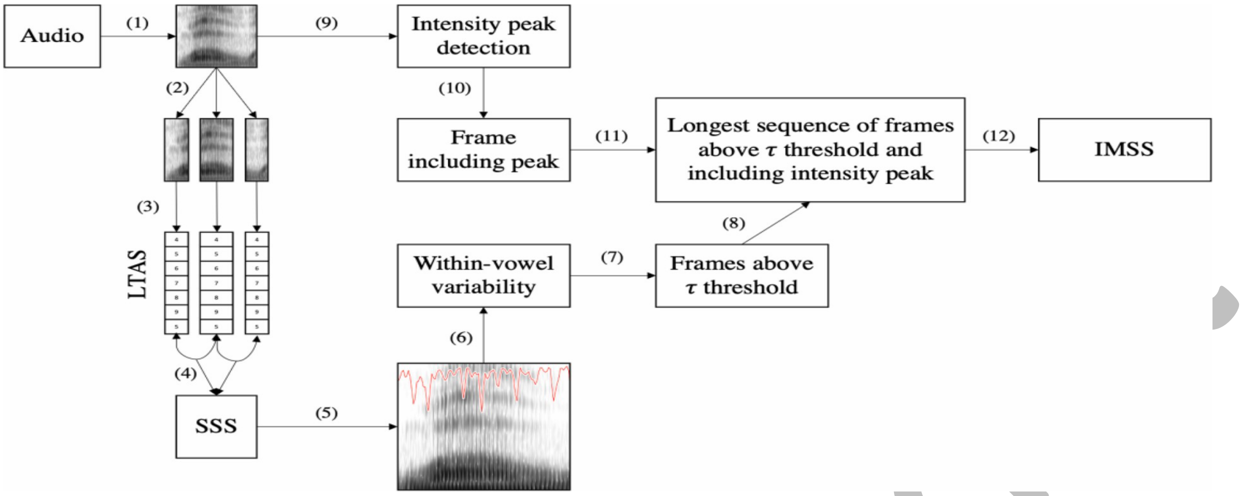


Fig. 1. Block diagram of the IMSS procedure.

However, it is important to note that it can also summarize spectral information over short periods of time. Specifically, it stands for the "logarithmic spectral density as a function of frequency, expressed in dB/Hz relative to 2 × 10–5 Pa" (Boersma and Weenink, 2021). It is interesting that the LTAS of a sound can be computed across the entire sampling frequency. This makes it possible to detect dissimilarity at higher frequencies than the algorithms which use formant tracking, similarly to metrics based on cepstral coefficients. Furthermore, the LTAS does not rely on a complex tracking algorithm: it only computes the logarithmic spectral density in a given frequency band averaged over time. If the sound is long enough,[2] the LTAS of any sound can be computed, unlike formant-based metrics which need an algorithm to detect the formants and it is well known that formant tracking is not always accurate, see Van der Harst (2011).

The LTAS makes it possible to obtain an overall representation of the speech as an *n*-dimensional vector which represents the logarithmic spectral density in each frequency bin (step (3) in Fig. 1). Therefore, it is possible to calculate the correlation between the LTAS of frames f and f - 1 as well as between f and f + 1 (step (4) in Fig. 1). Those values can be averaged to get a Spectral Stability Score which evaluates the similarity between frame f and the preceding and following frames, see Eq. (1) where SSS_f stands for the Spectral Stability Score of frame f, where n is the number of frequency bins, where x is frame f, where y is frame f - 1 and where z is frame f + 1. In the case of the first and last frames, the correlation is computed with the following and preceding frames, respectively, see Eqs. (1.a) and (1.b). The obtained scores range between 0 and 1. A score close to 0 indicates that successive frames are very dissimilar and this suggests spectral instability. Scores close to 1 indicate that adjacent frames are very similar and suggest spectral stability. Contrary to the previously mentioned metrics, the SSS is a stability metric, not an instability metric. The identified frames are the ones which compose the longest sequence of frames whose SSS is above the k^{th} within-vowel percentile of stability (step (7) in Fig. 1).

[2] Twenty ms is sufficient for frequency bands of 100 Hz in PRAAT (Boersma and Weesnink, 2021).

$$SSS_f = \left(\frac{\sum_{i=1}^{n}(x_i - \overline{x}) * (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})} * \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})}} + \frac{\sum_{i=1}^{n}(x_i - \overline{x}) * (z_i - \overline{z})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})} * \sqrt{\sum_{i=1}^{n}(z_i - \overline{z})}} \right) /2 \quad (1)$$

$$\mathrm{SSS}_{first-f} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) * (z_i - \overline{z})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})} * \sqrt{\sum_{i=1}^{n}(z_i - \overline{z})}} \quad (1.a.)$$

$$SSS_{last-f} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) * (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})} * \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})}} \quad (1.b.)$$

A potential limitation of this metric is that absolute silence is characterized by a perfect SSS. Therefore, the algorithm is implemented such that it identifies a portion of the acoustic signal only if it contains an intensity peak (steps (9–11) in Fig. 1). Duckworth et al. (2011: 40) recommend that measurements are made in a stable portion and around the intensity peak. However, Kent and Vorperian (2018: 77) suggest that the intensity peak does not match the spectrally most stable part of the vowel in some cases. As far as the intensity tracking is concerned, the standard parameters of PRAAT are used.[3] If intensity tracking cannot be applied, the longest sequence of frames whose SSS is above certain level of spectral stability is selected, not considering intensity.

4. The present work

The main aim of this paper is the experimental investigation of two methodological questions.

The first experiment investigates whether previously proposed spectral stability metrics differ in their ability to identify the *a priori* determined spectrally stable portion of synthesized vowels and how their efficiency compares to the conventional assumption that the spectrally stable portion is situated around the temporal center of a vowel. This was tested on a corpus of synthetic stimuli.

The second experiment investigates whether the metrics have an effect on acoustic measurements typically made by phoneticians – such as F0, F1 or F2 – by applying spectral stability metrics to the same corpus of synthesized speech.

The third experiment consists of the same analysis as the second experiment, but the analysis is carried out on a corpus of child speech that was elicited by means of a (non)word repetition task.

[3] Since the tracking of intensity requires that the physical duration of the sound be at least 6.4 divided by the minimum pitch, 100 Hz in our case, the intensity of the sounds whose duration is shorter than 0.064 s cannot be computed.

5. Experiment I - synthetic stimuli

Experiment I investigates the ability of 5 metrics to identify the spectrally stable portion of synthetic vowels whose spectrally stable portion had been experimentally varied. Their results were compared to the traditional assumption that the spectrally stable portion of a vowel is situated around a vowel's temporal midpoint.

5.1. Material and methods

In order to assess how well different metrics perform to identify the spectrally stable portion of vowel sounds, a range of synthetic speech stimuli were created in which spectral stability was carefully controlled. For this purpose, PRAAT's articulatory speech synthesis software was used (Boersma, 1998). This synthesizer can generate speech sounds by specifying the activity of 29 articulators by values ranging from -1 to +1. As such, the activity of the different articulators is modulated to create the shape of the pharyngeal, oral, and nasal tracts as well as the laryngeal and pulmonary activity. This synthesizer can also produce sequences of speech sounds by linearly interpolating the articulator activity values between two articulatory targets specified at two different timepoints (Boersma, 1998: 62).

In order to synthesize speech-like CVC sequences, the activity of the different articulators has to be specified at multiple timepoints throughout the sequence, i.e. each segment and each transition between segments is created by specifying the timepoints at which the articulators reach and depart from their targets. Consequently, the synthesis of a vowel can be designed so that all articulators have reached their target for the vowel and do not depart from it between two *a priori* determined timepoints. It means that none of the articulatory parameters varies between those two points, and it can therefore be assumed that this portion of the signal has the greatest possible stability within the synthetic CVC sequence. As such, the duration of the spectrally stable portion can be controlled. For maximizing the variability within the stimulus set, the duration of the spectrally stable portion of the vowels was generated randomly within certain pre-defined limits (see Section 5.1.1.). Similarly, the position of that stable portion, i.e. whether it is situated closer to one of the two consonants or perfectly in the middle, was also varied randomly.

As such, speech synthesis makes it possible to create CVC-like stimuli in which the duration and location of the steady-state portion of the vowel can be controlled as an experimental variable. It is important to emphasize that the aim of the simulation in this experiment is not to synthesize naturally sounding human speech, but rather to maximise the variation in the location and duration of the spectrally stable portion of the vowels in order to test how well the different metrics perform in detecting the experimentally controlled spectrally stable portion.

5.1.1. Stimuli

The stimuli for this experiment consisted of CVC-sequences in which an [a] vowel is preceded and followed by a single consonant from the following set [p, b, m, t, d, n, k, g, ŋ]. These sounds were synthesized by specifying 8 articulators of the synthesizer using the values given in Table 1 (a more formal characterisation can be found in the appendix B).

For the synthesis of CVC-sequences, articulators were specified for 12 points in time. Fig. 2 presents the gestural score of a hypothetical articulator with the 12 points at which its activity is determined for the synthesis of a given CVC-sequence (a more formal characterisation can be found in the appendix C). Fig. 3 shows the same articulatory trajectory and its implications in terms of coarticulation. Overall, the stimuli met three conditions:

1) Each sequence has a total duration of 350 ms;
2) Each articulator reaches its articulatory target for C1 and C2 at given timepoints;
3) The vowel has a spectrally stable portion of at least 50 ms.

Each CVC-sequence was generated by one of the three standard artificial speakers provided by the articulatory synthesizer: (i) a male speaker, (ii) a female speaker and (iii) a child. For each sequence, the 12 points in time at which the activity of the articulators was specified were generated randomly 20 times. This procedure resulted in 4860 synthesized CVC sequences (9 C1×1V×9 C2×3 speakers×20 timing sequences).

5.1.2. Evaluation of the metrics

The present paper compares 5 spectral stability identification procedures to a baseline condition (mid-third portion of the vowel, i.e. the most traditional identification method). The spectral stability procedures are listed in Table 2.

Time-defined		Mid-third IMSS
Feature-defined	Vowel-independent	Coefficient of change (cfr. Lennig, 1978) Slope in log F2-log F1space (cfr. Hillenbrand et al., 1995) Pooled within-variance (cfr. Van Bergem, 1988) Cepstral coefficients (cfr. Van Bergem, 1993) Spectral stability score
	Vowel-dependent	n/a

The different IMSS procedures were compared with respect to their ’precision’ and ’recall’. It is important that algorithms identify the relevant parts of the acoustic signal and exclude the irrelevant parts of it. Therefore, a score is needed which captures how well an algorithm classifies portions of the acoustic signal as stable or unstable. That task is very similar to the assessment of classifiers (e.g. Buckland and Gey, 1994) for two reasons. First, the amount of signal which is relevant should be maximized. The more of the relevant portion of the signal is selected, the better. In technical terms, it should maximize recall. Second, it should minimize the amount of signal, which is influenced by its phonetic environment, and which is thus not relevant, i.e. it should technically increase precision. Precision and recall can be calculated on the basis of Eqs. (2) and (3), respectively where TP stands for true positives (i.e. stable portion of the signal identified as stable), FP for false positives (i.e. unstable portion of the signal identified as stable) and FN for false negatives (i.e. stable portion of the signal identified as unstable). True/false positives/negatives are illustrated with respect to the synthesized CVC-sequence in Fig. 4.

$$precision = TP/(TP+FP) \qquad (2)$$

$$recall = TP/(TP+FN) \tag{3}$$

To obtain an overall measure of the accuracy of the algorithms in terms of both recall and precision, the f1-score was used, which is the harmonic average of precision and recall, see Eq. (4). This metric was chosen because it is a single score that evaluates both the precision and recall of an algorithm between 0 and 1. Higher f1-scores indicate a better trade-off between precision and recall. A score of 0 indicates that an algorithm completely failed to locate the stable portion of the vowel. A score of 1 indicates that the algorithm perfectly detected the spectrally stable portion.[4] For the analysis, the f1-scores were transformed to their logits. In conclusion, higher logits of f1 indicate better identification of the spectrally stable vowel portion.

$$\mathrm{f1} = 2 * \frac{\mathrm{precision} * \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{2}$$

5.1.3. Computational implementation of the evaluation procedure

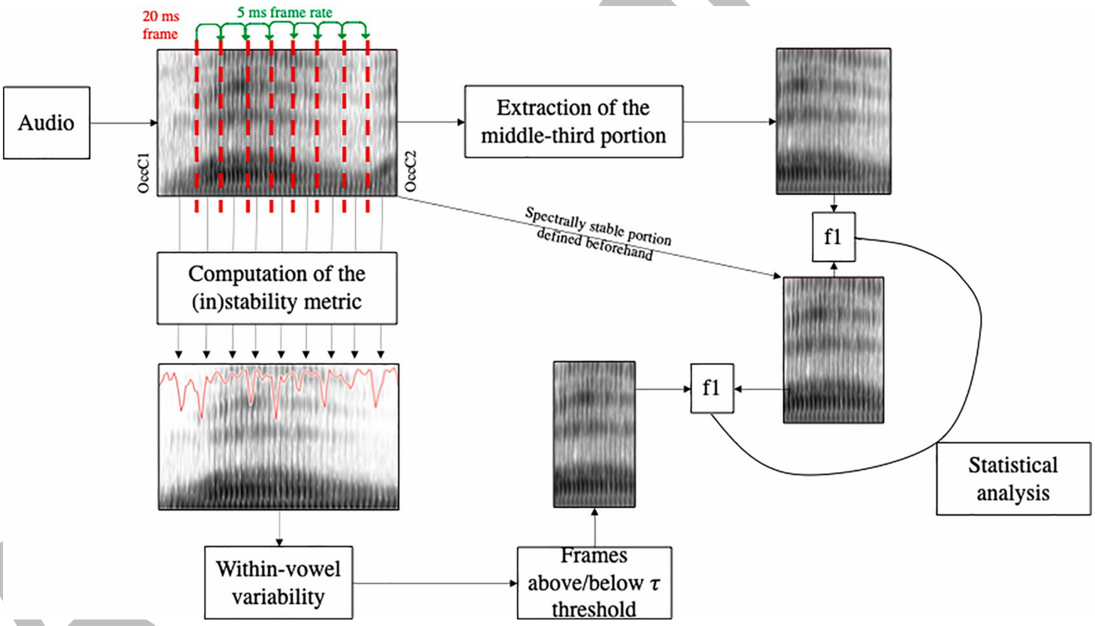


Fig. 5. Block diagram of the evaluation procedure.

The overall evaluation procedure is summarized in Fig. 5. The metrics were coded in algorithms which take the form of Python scripts.[5] Features were calculated by means of the Parselmouth API (Jadoul et al., 2018) of PRAAT (Boersma and Weenink, 2021). Each stimulus was segmented in overlapping frames of 20 ms with a frame rate of 5 ms. The

[4] A problem arises when the denominator in Eq. (4) is equal to 0 but for it to be equal to 0, both the precision and recall need to be equal to 0. This would mean that the algorithm has no precision and no recall. Therefore, it is reasonable to assign to the algorithm a very low f1 value of $1{*}10^{10}$. [5] The scripts are available upon simple request to the authors.

algorithm then computed the (in)stability metric for each frame and selected the spectrally most stable frames, i.e. the frames whose stability is above a given threshold value.

Phone	*Lungs*	*Interarytenoid*	*LevatorPalatini*	*Masseter*	*Hyoglossus*	*UpperTongue*	*Styloglossus*	*OrbicularisOris*
[a]	0	0.50	1	-0.5	0.5	0	0	0
[b]	0	0.53	1	0.5	0.0	0	0	1
[p]	0	0.00	1	0.5	0.0	0	0	1
[m]	0	0.53	0	0.5	0.0	0	0	1
[d]	0	0.53	1	0.0	0.0	1	0	0
[t]	0	0.00	1	0.0	0.0	1	0	0
[n]	0	0.53	0	0.0	0.0	1	0	0
[g]	0	0.53	1	0.0	0.0	0	1	0
[k]	0	0.00	1	0.0	0.0	0	1	0
[ŋ]	0	0.53	0	0.0	0.0	0	1	0
Neutral position	0	0.00	0	0.0	0.0	0	0	0

Table 1. Articulatory parameters for the targets of each phone used in the simulation. The articulatory target specifications are adapted from Boersma (1998) and the PRAAT scripts available on its companion website.

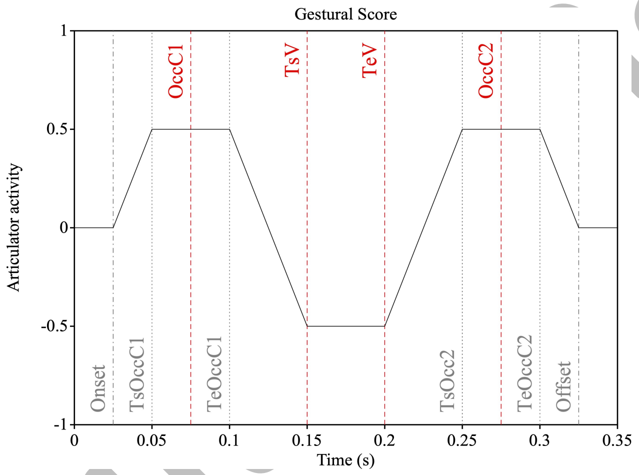


Fig. 2. Gestural score of a hypothetical articulator with the different time- points needed for the synthesis of the stimuli.

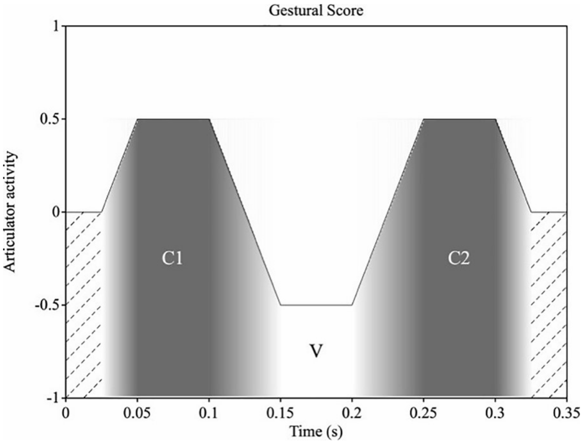


Fig. 3. Model of coarticulation for a synthesized CVC sequence; gray: target of C reached, white: target of V reached, hatched: neutral position.

In order to assess the ability of different metrics to capture spectral stability, they need to be compared on the same basis, i.e. the threshold should be optimized for each metric. For that purpose, the overall within-stimulus stability is evaluated and rescaled to 1. As such, vowel portions which reach at least k^{th}percentile of stability (for stability metrics) or which are below k^{th} percentile of stability (for instability metrics) can be selected for IMSS. By making k vary on a continuum from 0 to 1, a maximum or minimum of variability can be allowed. As such, the k^{th} percentile with which the metric leads to better IMSS can be observed. The k^{th} percentile of (in)stability within-stimulus is henceforth

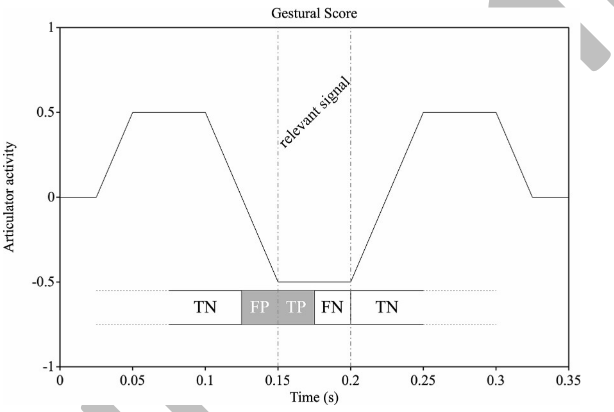


Fig. 4. Gestural score of a hypothetical articulator; gray area: potential selection by an algorithm. referred to as the τ threshold.

In other words, the algorithms select all sequences of at least 2 consecutive frames whose stability score is above τ or the frames whose instability score is below τ. Among all those potential sequences, the algorithm then selects the longest possible one.

5.1.4. Statistical analysis

The statistical analysis was carried out in R (Development Core Team, 2021) and the R package *lme4* (Bates et al., 2015). The package *lmerTest* (Kuznetsova et al., 2015) was used to obtain *p*-values. Multilevel modelling was used to determine the potential effect of the chosen technique via pairwise comparisons between the IMSS on the mid-third part of the vowel and the IMSS by one of the feature-defined algorithms whose τ threshold had been set at a particular value. The dependent variable of the model is the logit of the f1-score. The fixed effects are the voicing of C1 and of C2, the duration of the *a priori* determined stable portion, the duration of the transition phase between the C1 constriction and the start of the stable portion, and the duration of the transition between the end of the stable portion and the C2 constriction (without interaction). Besides that, C1 and C2, as well as speakers (i.e. male, female or child) and items (i.e. each unique combination of C1 and C2), were included as random intercepts.

The duration of the transition between C1 and V was calculated as the difference between the earliest point at which at least one articulator starts to depart from its C1 target towards its V target and the moment at which all articulators have reached their V target divided by the length of the sound between the complete constriction of C1 and C2. The duration of the transition between V and C2 was calculated as the difference between the end of the stable portion of the vowel and the point at which the last articulator reached its C2 target divided by the length of the sound between the complete constriction of C1 and C2.

The voicing of C1 and C2 were also included as fixed effects in the model because it is expected that voiced consonants result in smoother transitions than voiceless consonants. By including that effect in the model, it can be taken into account how the algorithms react to these constraints.

5.1.5. Threshold selection

Prior to the final analysis, information about how the different algorithms perform according to the amount of variability tolerated in the selection was collected by making τ vary from a minimal to a maximal value. It is expected that the accuracy of an algorithm (evaluated by its estimate with respect to IMSS on the mid-third) might behave (curvi) linearly, according to the amount of (in)variability tolerated, i.e. according to the chosen τ. To reduce the computational cost and observe the behavior of τ values according to the different metrics, 100 randomly selected stimuli were extracted from the full synthesized corpus. The algorithms were then run with different τ thresholds, ranging from 0.025 to 0.975 by steps of 0.025 on the subset of the corpus. We aimed to select four τ values which get the best results in order to test the metrics on the entire corpus with those specific τ values. It was expected to find some relation between the τ values, i.e. the amount of variability tolerated, and the quality of the output of the algorithm. The observed trends were then used to test the algorithms with the τ which potentially provide the best results.

If the algorithms perform best towards one or the other end of the τ continuum, the retained τ values are the four most extreme values towards that end. If the relation between the output of the algorithm and τ shapes a negative parabola, the τ value which gives the best results, i.e. its vertex, and the three lower surrounding τ values are retained. If the parabola opens upwards, the retained τ values are the two highest points as well as the preceding or following τ value which gets the lowest intercept. By doing so, the τ parameter which gives the best possible output per metric can be empirically determined and consequently, it can be avoided that a human decision plays a role.

5.2. Results

5.2.1. Selection

For each algorithm, the retained τ values are the thresholds with which the algorithm has the best possible output when compared to the traditional IMSS on the mid-third. The results are presented in Figs. 6–10. A positive intercept indicates that the algorithm performs better in IMSS than the traditional IMSS on the mid-third. A negative intercept indicates that IMSS on the mid-third is better than the one suggested by the algorithm. If the difference between both IMSS techniques is significant, the intercept is indicated by a red dot. Black dots indicate that the difference is not significant. These values were used to run the algorithm on the full corpus of synthetic stimuli.

As far as Lennig's (1978) coefficient of change is concerned, a positive parabolic-like relation is observed. It probably indicates that high τ values allow for very little variability within the IMSS. Thus, the algorithm selects relevant portions of the acoustic signal but probably does not maximize the amount of relevant signal in the selection. It probably indicates good precision, but inferior recall. On the contrary, lower τ values probably allow for more variability in the selection and improve recall but might consequently select a portion of the acoustic signal that is not relevant. This is probably why the algorithms perform best at both ends of the τ value continuum, as can be seen in Fig. 6. The τ values retained for further investigation were: 0.05, 0.075, 0.95, 0.975.

When it comes to the metrics of the slope in the log F2 – log F1 space, it can be seen in Fig. 7 that τ threshold values between 0.65 and 0.725 lead to better results. This may indicate that higher τ threshold values increase precision and lower values increase recall. Overall, intermediate values lead to better f1 scores.[5] The retained τ values were 0.65, 0.675, 0.7 and 0.725.

With respect to the pooled within-variance of the log-transformed formant values, τ threshold values around 0.5 lead to better results which is clear from the parabolic-like relation between τ and the estimate displayed in Fig. 8. This may indicate that higher τ threshold values increase precision and lower τ increase recall. The retained τ values for further analysis were: 0.475, 0.5, 0.525, 0.55.

When it comes to the metrics of cepstral coefficients, it is clear from Fig. 9 that higher τ threshold values lead to better results than lower τ values. This means that the results are better when less instability is accepted by the algorithm. The values which will therefore be retained were: 0.975, 0.95, 0.925 and 0.9.[6]

As far as the SSS algorithm is concerned, Fig. 10 suggests that lower τ thresholds provide better accuracy of the algorithm with respect to the traditional IMSS on the mid-third. It indicates that the algorithm with less strict τ values or which allow for more spectral variability in the selection results in selections which are closer to the true spectrally stable portion. Overall, the algorithm does not seem to perform better than the traditional IMSS

[5] τ=0.825 and τ=0.875 are not represented because the models do not converge.

[6] τ=0.025 is not represented because none of the 100 stimuli could be processed with that parameter. This is probably the result of the very low level of accepted instability associated to this extreme τ value.

method, except for the lowest τ value tested, i.e. 0.025, but the difference did not reach significance. Therefore, the four retained τ threshold values were: 0.025, 0.05, 0.075 and 0.1.

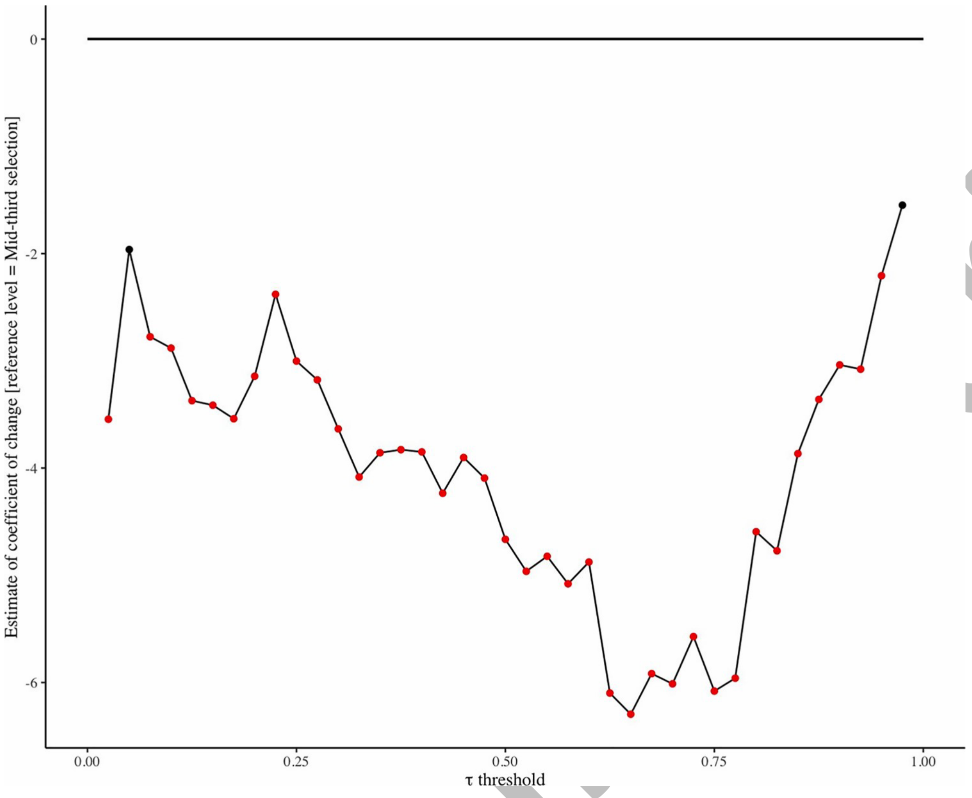


Fig. 6. Estimate of the fixed effect of metric (coefficient of change) [reference level = Mid-third IMSS]. Red dot: significant effect ($p < 0.05$); black dot: not significant effect.

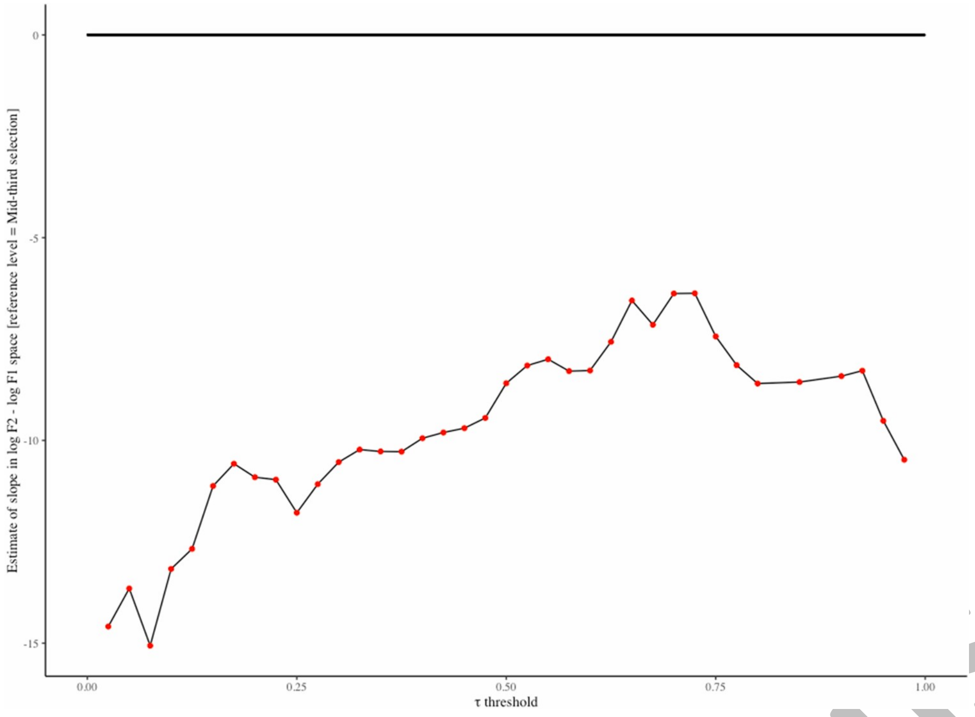


Fig. 7. Estimate of the fixed effect of metric (slope in log F2 - log F1 space) [reference level = Mid-third IMSS]. Red dot: significant effect (p < 0.05); black dot: not significant effect.

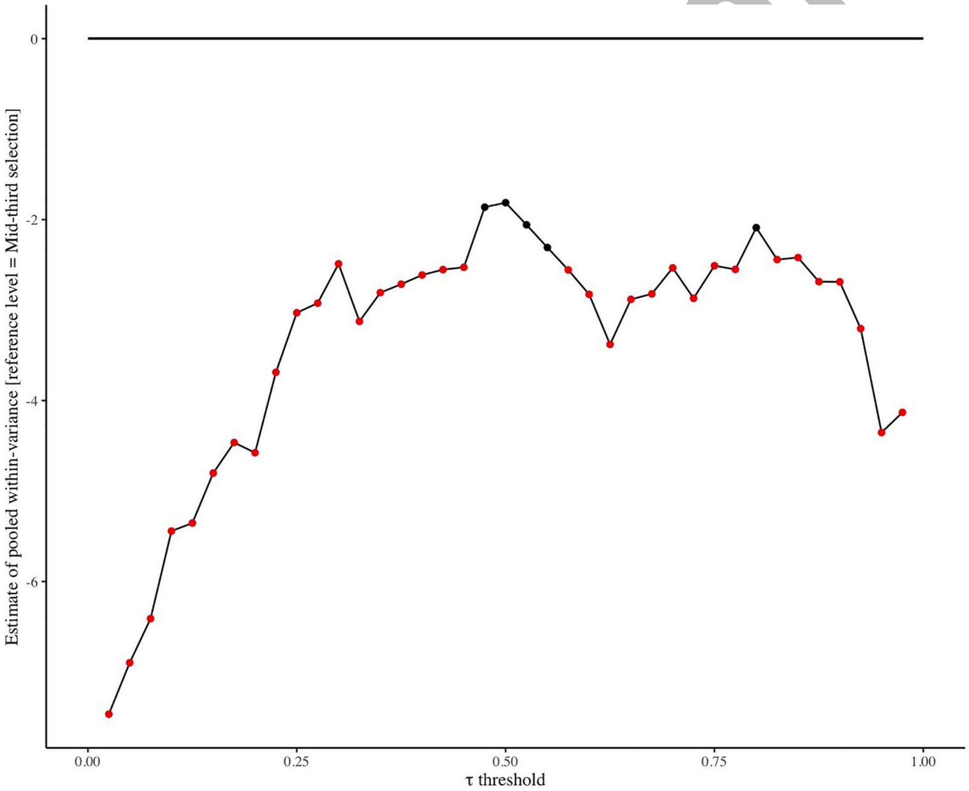


Fig. 8. Estimate of the fixed effect of metric (pooled within-variance) [reference level = Mid-third IMSS]. Red dot: significant effect (p < 0.05); black dot: not significant effect.

5.3. Discussion pooled within-variance of the log transformed values is significant, but the effect is negative. This indicates that the IMSS around the middle of

The results of this experiment indicate that the effect of the metric for the vowel performs better because it has a higher logit of f1-score than the coefficient of change, the slope in log F2 - log F1 space and the the feature-defined IMSS.

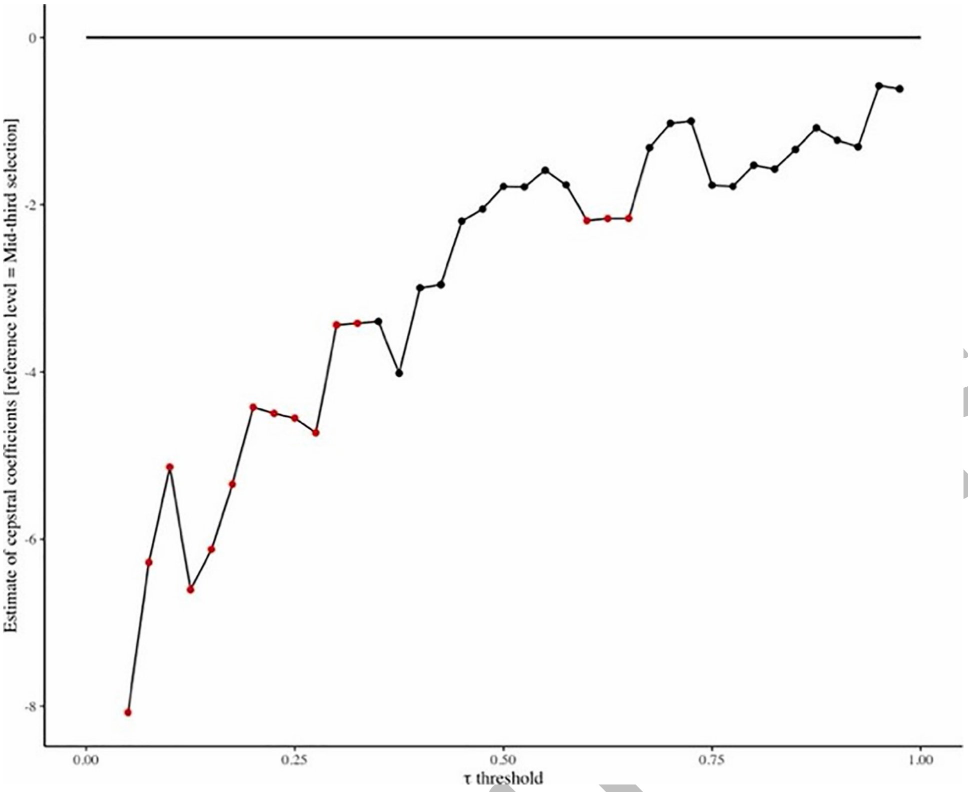


Fig. 9. Estimate of the fixed effect of metric (cepstral coefficients) [reference level = Mid-third IMSS]. Red dot: significant effect ($p < 0.05$); black dot: not significant effect.

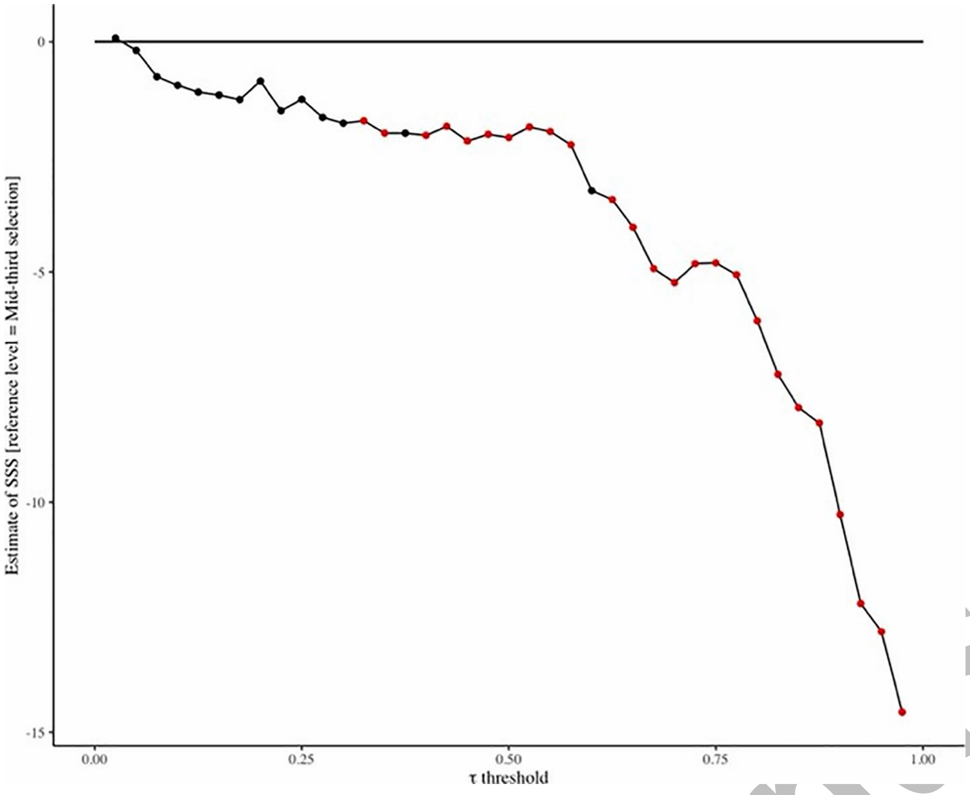


Fig. 10. Estimate of the fixed effect of metric (SSS) [reference level = Mid-third IMSS]. Red dot: significant effect (p < 0.05); black dot: not significant effect.

5.2.2. Full analysis

The identification procedures based on the different metrics have been applied to the corpus of 4860 synthetic CVC sequences, with the τ values expected to give the best results according to the threshold selection test. Table 3 summarizes the results of the fixed effects of the linear mixed-modelling procedure for each metric and each selected τ.

The estimate of the Algorithm factor indicates whether the algorithm gives better results, i.e. higher logit of f1-scores than the traditional IMSS on the mid-third. Those estimates are printed in bold in Table 3. A positive estimate shows that using the feature-defined algorithm rather than the traditional time-defined algorithm increases the logit of the f1- score. A negative estimate suggests that the traditional method performs best. The *p*-values give the significance level of the effect.

From Table 3, it is clear that the results vary largely according to the metric and the τ value used. One parameter, however, consistently affects the performance of all algorithms, i.e. the duration of the vowel. Longer vowels facilitate the identification of the spectrally stable vowel portion. A similar, but less consistent effect (across metrics and τ) of the transition durations 1 and 2 is also found. Longer transitions are less abrupt in terms of their frame-by-frame stability and the more difficult it is for the algorithm to detect spectrally stable portions. The output of some algorithms can also be influenced by the voicing of the C2. There can be a significant effect which shows that a voiced C2 is slightly easier for the algorithm to process. The opposite was expected because voiced consonants trigger smoother transitions. This unexpected pattern may result from the constriction of voiceless consonants associated with a lower amount of energy in the spectrum. As a result, the transition patterns of the articulators may not be reflected as clearly in the spectrum as in a voiced consonant in which transition patterns are more easily identified because of the increased energy present in the signal. However, the effect of C1 voicing is never significant.

Concerning the metrics themselves, it can be observed that there are significant differences between the tested algorithms and the traditional identification of the stable portion at the center of the vowel. Most results indicate that the selection around the middle of the vowel outperforms the selection of a feature-defined algorithm. This holds for the coefficient of change, the slope in log F2 - log F1 space and pooled within-variance metric. However, the algorithm using cepstral coefficients as a metric does not do significantly worse than the IMSS around the center of the vowel. The 0.975 τ value in fact gives even significantly better results for the feature-based algorithm than for the traditional selection. Similarly, the algorithm based on the SSS leads to significantly better results with all 4 selected τ values.

Table 3. Results of the fixed effects as a function of metric and τ value.

Algorithm and τ value	Estimate	SE	df	*t*-value	*p*-value
Coefficient of change - 0.025					
(Intercept)	0.68629	0.37386	163.06169	1.836	0.0682.
Voicing C1	0.04095	0.18262	6.98992	0.224	0.8290
Voicing C2	0.38320	0.14846	6.91653	2.581	0.0368*
Duration stable portion	11.04906	1.76830	2564.17457	6.248	4.84e-10***
Duration transition 1	-39.10746	1.45052	2908.43492	-26.961	<2e-16***
Duration transition 2	-1.29275	1.85591	3267.65868	-0.697	0.4861
Algorithm [Mid-third IMSS]	**-1.58961**	**0.23298**	**347.54130**	**-6.823**	**3.98e-11*** **
Coefficient of change - 0.95					
(Intercept)	-1.9342	0.4642	23.1282	-4.167	0.000368***
Voicing C1	-0.2499	0.4157	7.0001	-0.601	0.566759
Voicing C2	-0.1070	0.1612	6.9808	-0.664	0.527914
Duration stable portion	29.7523	1.6961	6247.1783	17.542	<2e-16***
Duration transition 1	-34.1396	1.3721	7281.8454	-24.882	<2e-16***
Duration transition 2	16.7725	1.8508	8004.7332	9.063	<2e-16***
Algorithm [Mid-third IMSS]	**-2.3797**	**0.1568**	**333.3879**	**-15.172**	**2e-16*** **
Slope in log F2 - log F1 space - 0.65					
(Intercept)	-4.7326	0.9657	11.5677	-4.901	0.000407***
Voicing C1	0.7138	0.8311	7.0008	0.859	0.418820
Voicing C2	0.2195	0.3162	6.9869	0.694	0.509891
Duration stable portion	51.3916	2.0121	7478.7223	25.541	<2e-16***
Duration transition 1	-41.0956	1.6232	8075.3380	-25.318	<2e-16***
Duration transition 2	22.2936	2.1791	8587.9394	10.231	<2e-16***
Algorithm [Mid-third IMSS]	**-7.8153**	**0.2071**	**548.9717**	**-37.738**	**<2e-16*** **
Slope in log F2 - log F1 space - 0.7					
(Intercept)	-4.7721	0.8686	11.4954	-5.494	0.00016***
Voicing C1	0.8107	0.7461	7.0021	1.087	0.31319
Voicing C2	0.1584	0.2222	6.9781	0.713	0.49907
Duration stable portion	52.3589	1.9970	7613.8713	26.219	<2e-16***
Duration transition 1	-42.6086	1.6107	8166.2892	-26.454	<2e-16***
Duration transition 2	23.5459	2.1617	8653.2565	10.892	<2e-16***
Algorithm [Mid-third IMSS]	**-7.5883**	**0.2073**	**586.1858**	**-36.605**	**<2e-16*** **
Pooled within-variance 0.475					
(Intercept)	-2.7424	0.4256	9.9635	-6.444	7.53e-05***
Voicing C1	-0.1919	0.1494	7.0174	-1.284	0.23982
Voicing C2	0.7675	0.1735	6.9959	4.424	0.00307**
Duration stable portion	33.8571	1.6082	6785.8019	21.053	<2e-16***
Duration transition 1	-32.9222	1.2990	7633.3442	-25.344	<2e-16***
Duration transition 2	9.7132	1.7485	8280.3093	5.555	2.86e-08***
Algorithm [Mid-third IMSS]	**-2.0465**	**0.1565**	**395.2295**	**-13.077**	**<2e-16*** **
Pooled within-variance 0.525					
(Intercept)	-2.8859	0.3738	17.2419	-7.720	5.38e-07***
Voicing C1	-0.2316	0.1494	7.0170	-1.550	0.16491
Voicing C2	0.7431	0.1452	6.9914	5.117	0.00138**
Duration stable portion	33.6234	1.5921	6674.8947	21.118	<2e-16***
Duration transition 1	-31.1211	1.2861	7557.0732	-24.197	<2e-16***
Duration transition 2	10.9098	1.7322	8207.6175	6.298	3.17e-10***
Algorithm [Mid-third IMSS]	**-1.8274**	**0.1541**	**374.8195**	**-11.860**	**<2e-16*** **
Cepstral coefficients - 0.9					
(Intercept)	-0.4422	0.3640	63.7945	-1.215	0.2290
Voicing C1	0.1700	0.2549	6.9805	0.667	0.5261
Voicing C2	0.3704	0.1235	8814.0832	3.001	0.0027**
Duration stable portion	23.9209	1.5692	5144.6626	15.244	<2e-16***
Duration transition 1	-48.8692	1.2955	5811.8946	-37.721	<2e-16***
Duration transition 2	10.3963	1.6813	6392.5504	6.183	6.66e-10***
Algorithm [Mid-third IMSS]	**-0.3548**	**0.1440**	**518.9475**	**-2.464**	**0.0141***
Cepstral coefficients - 0.95 †					
(Intercept)	-0.4095	0.3120	148.5770	-1.312	0.19138
Voicing C1	0.2328	0.1742	6.9764	1.336	0.22348
Voicing C2	0.3490	0.1163	8810.2721	3.001	0.00269**
Duration stable portion	20.4515	1.4705	4835.2594	13.908	<2e-16***
Duration transition 1	-43.5009	1.2147	5577.9263	-35.811	<2e-16***
Duration transition 2	7.1854	1.5772	6157.3569	4.556	5.32e-06***
Algorithm [Mid-third IMSS]	**0.2048**	**0.1328**	**463.0645**	**1.542**	**0.12364**
SSS - 0.025					
(Intercept)	-0.3991	0.2621	76.9105	-1.523	0.1319
Voicing C1	-0.1347	0.1199	7.0058	-1.124	0.2980
Voicing C2	0.3703	0.1642	7.0013	2.255	0.0588.
Duration stable portion	10.4954	1.1585	7007.1928	9.059	<2e-16***
Duration transition 1	-23.6886	0.9343	7700.1546	-25.355	<2e-16***
Duration transition 2	-1.8632	1.2572	8346.3640	-1.482	0.1384
Algorithm [Mid-third IMSS]	**1.2209**	**0.1183**	**405.4763**	**10.317**	**<2e-16*** **
SSS - 0.075					
(Intercept)	-0.6403	0.2791	77.6245	-2.295	0.0245*
Voicing C1	-0.1730	0.1535	7.0093	-1.127	0.2970
Voicing C2	0.5554	0.1589	7.0022	3.495	0.0101*
Duration stable portion	10.2381	1.2347	6634.9612	8.292	<2e-16***
Duration transition 1	-21.2268	0.9964	7440.5098	-21.304	<2e-16***
Duration transition 2	-1.9502	1.3417	8154.6348	-1.454	0.1461
Algorithm [Mid-third IMSS]	**0.6514**	**0.1230**	**350.7009**	**5.296**	**2.1e-07*** **
Algorithm and τ value	Estimate	SE	df	*t*-value	*p*-value
Coefficient of change - 0.05					
(Intercept)	-0.61862	0.32886	80.69081	-1.881	0.0636.
Voicing C1	-0.08622	0.12925	6.92273	-0.667	0.5263
Voicing C2	0.56765	0.20907	6.98569	2.715	0.0300*
Duration stable portion	16.18396	1.50487	4361.78891	10.754	<2e-16***
Duration transition 1	-29.47527	1.21671	5197.47244	-24.225	<2e-16***
Duration transition 2	-1.28751	1.59964	5717.22484	-0.805	0.4209
Algorithm [Mid-third IMSS]	**-2.24912**	**0.13727**	**471.03038**	**-16.385**	**<2e-16*** **
Coefficient of change - 0.975					
(Intercept)	-1.92312	0.37794	28.60736	-5.088	2.06e-05***
Voicing C1	-0.07625	0.31361	7.00228	-0.243	0.815
Voicing C2	0.18629	0.12154	9563.09101	1.533	0.125
Duration stable portion	26.43563	1.45880	6354.51542	18.121	<2e-16***
Duration transition 1	-33.25841	1.17932	7329.77693	-28.201	<2e-16***
Duration transition 2	15.27435	1.58973	8052.54385	9.608	<2e-16***
Algorithm [Mid-third IMSS]	**-0.59887**	**0.13776**	**335.56019**	**-4.347**	**1.83e-05*** **
Slope in log F2 - log F1 space - 0.675					
(Intercept)	-4.7926	0.9181	11.5514	-5.220	0.000243***
Voicing C1	0.7608	0.7894	7.0013	0.964	0.367275
Voicing C2	0.2168	0.2727	6.9832	0.795	0.452664
Duration stable portion	51.8769	2.0064	7588.6654	25.855	<2e-16***
Duration transition 1	-41.8996	1.6184	8150.6107	-25.890	<2e-16***
Duration transition 2	23.5415	2.1722	8642.5044	10.838	<2e-16***
Algorithm [Mid-third IMSS]	**-7.7161**	**0.2079**	**579.6101**	**-37.119**	**<2e-16*** **
Slope in log F2 - log F1 space - 0.725					
(Intercept)	-4.6580	0.8177	12.4931	-5.696	8.55e-05***
Voicing C1	0.7040	0.7129	7.0015	0.988	0.356
Voicing C2	0.2346	0.2027	6.9713	1.157	0.285
Duration stable portion	51.6113	1.9977	7457.6246	25.835	<2e-16***
Duration transition 1	-42.6823	1.6118	8063.7346	-26.481	<2e-16***
Duration transition 2	23.3664	2.1638	8578.2382	10.799	<2e-16***
Algorithm [Mid-third IMSS]	**-7.5786**	**0.2047**	**543.8825**	**-37.018**	**<2e-16*** **
Pooled within-variance 0.5 †					
(Intercept)	-2.7478	0.4040	9.9346	-6.802	4.89e-05***
Voicing C1	-0.1589	0.1327	9564.4283	-1.198	0.23109
Voicing C2	0.7133	0.1408	6.9908	5.065	0.00146**
Duration stable portion	33.4369	1.6006	6749.3648	20.891	<2e-16***
Duration transition 1	-32.5110	1.2928	7607.4523	-25.149	<2e-16***
Duration transition 2	10.0361	1.7407	8245.6061	5.765	8.44e-09***
Algorithm [Mid-third IMSS]	**-1.9136**	**0.1557**	**384.9238**	**-12.291**	**<2e-16*** **
Pooled within-variance 0.575					
(Intercept)	-2.9854	0.3438	25.2424	-8.682	4.72e-09***
Voicing C1	-0.2546	0.1305	9568.1126	-1.951	0.0511.
Voicing C2	0.7545	0.1304	9566.7181	5.786	7.44e-09***
Duration stable portion	33.1043	1.5693	6494.7343	21.094	<2e-16***
Duration transition 1	-29.5468	1.2682	7434.2518	-23.298	<2e-16***
Duration transition 2	11.8676	1.7083	8117.4922	6.947	4.02e-12***
Algorithm [Mid-third IMSS]	**-1.6380**	**0.1502**	**351.0263**	**-10.906**	**<2e-16*** **
Cepstral coefficients - 0.925					
(Intercept)	-0.4415	0.3319	127.0916	-1.330	0.18584
Voicing C1	0.1765	0.1935	6.9684	0.912	0.39219
Voicing C2	0.3802	0.1216	8811.1290	3.128	0.00177**
Duration stable portion	22.7304	1.5405	4949.1122	14.755	<2e-16***
Duration transition 1	-46.8129	1.2723	5661.4977	-36.795	<2e-16***
Duration transition 2	9.0615	1.6516	6243.8313	5.487	4.26e-08***
Algorithm [Mid-third IMSS]	**-0.1732**	**0.1401**	**482.0791**	**-1.236**	**0.21689**
Cepstral coefficients - 0.975					
(Intercept)	-0.75285	0.29469	165.71362	-2.555	0.0115*
Voicing C1	0.08821	0.13680	6.99516	0.645	0.5396
Voicing C2	0.43488	0.13234	7.07073	3.286	0.0132*
Duration stable portion	20.04261	1.41276	4753.00323	14.187	<2e-16***
Duration transition 1	-37.58980	1.16713	5511.85107	-32.207	<2e-16***
Duration transition 2	6.19361	1.51564	6089.90480	4.086	4.44e-05***
Algorithm [Mid-third IMSS]	**0.54407**	**0.12726**	**448.85082**	**4.275**	**2.33e-05*** **
SSS - 0.05					
(Intercept)	-0.5766	0.2759	67.1756	-2.090	0.0404*
Voicing C1	-0.1195	0.1375	7.0061	-0.869	0.4137
Voicing C2	0.4821	0.1763	6.9997	2.735	0.0291*
Duration stable portion	9.9587	1.1893	6837.3385	8.373	<2e-16***
Duration transition 1	-22.1793	0.9594	7579.7769	-23.118	<2e-16***
Duration transition 2	-1.1046	1.2914	8258.4438	-0.855	0.3924
Algorithm [Mid-third IMSS]	**0.9649**	**0.1202**	**378.3950**	**8.025**	**1.29e-14*** **
SSS - 0.1					
(Intercept)	-0.8405	0.2923	68.8638	-2.876	0.005356**
Voicing C1	-0.1889	0.1417	7.0096	-1.333	0.224141
Voicing C2	0.7084	0.1870	7.0012	3.788	0.006816**
Duration stable portion	10.7218	1.2757	6757.3338	8.405	<2e-16***
Duration transition 1	-19.9152	1.0293	7530.6462	-19.348	<2e-16***
Duration transition 2	-3.0988	1.3858	8221.4365	-2.236	0.025370*
Algorithm [Mid-third IMSS]	**0.4446**	**0.1279**	**368.4903**	**3.477**	**0.000567*** **

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’. †: The model does not converge with the “nloptwrap” optimizer, and the data is therefore modelled with the “bobyqa” optimizer.

The metric based on cepstral coefficients was overall not significant is one τ value with which the IMSS procedure based on the cepstral when compared to the traditional method, which means that both coefficients gives statistically significant better results. Given that this techniques perform equally well: they capture a similarly good ratio of significant positive result is tied to that very specific τ, it is difficult to relevant and irrelevant portions of the speech signal. Nevertheless, there establish whether the significance of the result can be attributed to a parameter that would make the performance of the metric significantly better in general or to a τ especially suitable for the present corpus of synthesized stimuli.

The SSS metric performs significantly better than the traditional method with all of the four tested τ. This indicates that using this IMSS method with those specific parameters enables better identification of the spectrally stable portion than the IMSS around the temporal middle region of the vowel. Thus, measurements obtained after applying SSS could be more reliable than those obtained via the traditional approach.

6. Experiment II – synthetic stimuli

Experiment II studies to what extent the IMSS by the different metrics correlate with the mid-third IMSS on the mid-third in terms of starting point, center and end point, but most importantly, in terms of standard acoustic measurements made by phoneticians such as F0, F1, or F2. This experiment is designed to observe whether the ultimate goal of the phonetician, i.e. the measurements are considerably affected by adopting one or another technique, despite differences in the quality of the IMSS.

6.1. Materials and methods

6.1.1. Stimuli

The same data as in Experiment I were used.

6.1.2. Acoustic analysis

The F0, F1 and F2 of all vowels were analysed by means of a Python script on the portion of the vowel identified as stable by the different metrics. The measurement of F0, F1 and F2 was carried out by the Parselmouth API (Jadoul et al., 2018) of PRAAT (Boersma and Weenink, 2021) via its standard auto-correlation algorithm. The pitch floor and ceiling were set to 75 and 600 Hz, respectively. The maximum number of candidates was set to 15, the silence threshold to 0.03, the voicing threshold to 0.45, the octave cost to 0.01, the octave-jump cost to 0.35, the voiced/unvoiced cost to 0.14. As far as the formant parameters are concerned, the maximum number of formants was set to 5, the maximum formant to 5500 Hz, the window length to 0.025 and the pre-emphasis to 50. The F0, F1 and F2 values of each vowel were measured as the mean of all the measurements inside the selection determined by the different metrics.

6.1.3. Statistical analysis

On the one hand, it can be assessed to what extent the different algorithms correlate in terms of starting, middle and end points. On the other hand, it can be established whether the acoustic measurements carried out on the segmented audio signals differ from those obtained in the middle third portion of the vowels. To this end, F0, F1 and F2 in Hz were measured in each vowel. To compare the different outputs, Pearson's correlation coefficients were used with a significance level set at 0.05.

In addition to those two parameters, it is of interest to observe to what extent it is computationally feasible to apply the different metrics, i.e. the proportion of stimuli for which a given metric is able to identify a stable portion. It is essential for phoneticians to know whether an algorithm can be consistently applied to the majority of the analysed stimuli rather than only a small subset. This evaluation criterion, referred to as coverage in the remainder of this paper, is computed by dividing the number of processed stimuli by the total number of stimuli.

6.2. Results

Table 4 displays the correlation coefficients between the acoustic measurements (F0, F1 and F2) carried out in the spectrally stable portion of the vowels as determined by the different metrics and the same acoustic measurements made around the temporal center of the vowels. Table 4 also shows the correlation coefficients between the different metrics and the IMSS on the mid-third in terms of starting point, center and ending. In addition, Table 4 presents the coverage for each selected metric and τ value.

Algorithm	τ	Starting point		Centre		End point		F0		F1		F2		Coverage
		r	*p*-value	r	*p*-value	r	*p*-value	r	*p*-value	r	*p*-value	r	*p*-value	
Coefficient of change	0.05	0.564	< 2.22e-16	0.502	< 2.22e-16	0.578	< 2.22e-16	0.857	< 2.22e-16	0.515	< 2.22e-16	0.608	< 2.22e-16	0.724
	0.075	0.571	< 2.22e-16	0.466	< 2.22e-16	0.592	< 2.22e-16	0.855	< 2.22e-16	0.513	< 2.22e-16	0.608	< 2.22e-16	0.885
	0.95	0.444	< 2.22e-16	0.65	< 2.22e-16	0.762	< 2.22e-16	0.854	< 2.22e-16	0.56	< 2.22e-16	0.643	< 2.22e-16	0.998
	0.975	0.349	< 2.22e-16	0.784	< 2.22e-16	0.795	< 2.22e-16	0.904	< 2.22e-16	0.673	< 2.22e-16	0.762	< 2.22e-16	0.998
Slope in log F2 – log F1 space	0.65	0.314	< 2.22e-16	0.576	< 2.22e-16	0.413	< 2.22e-16	0.713	< 2.22e-16	0.377	< 2.22e-16	0.448	< 2.22e-16	0.998
	0.675	0.312	< 2.22e-16	0.588	< 2.22e-16	0.415	< 2.22e-16	0.713	< 2.22e-16	0.381	< 2.22e-16	0.46	< 2.22e-16	0.998
	0.7	0.321	< 2.22e-16	0.598	< 2.22e-16	0.425	< 2.22e-16	0.714	< 2.22e-16	0.398	< 2.22e-16	0.479	< 2.22e-16	0.998
	0.725	0.333	< 2.22e-16	0.604	< 2.22e-16	0.438	< 2.22e-16	0.709	< 2.22e-16	0.406	< 2.22e-16	0.486	< 2.22e-16	0.998
Pooled within-variance	0.475	0.544	< 2.22e-16	0.603	< 2.22e-16	0.631	< 2.22e-16	0.847	< 2.22e-16	0.475	< 2.22e-16	0.589	< 2.22e-16	0.998
	0.5	0.556	< 2.22e-16	0.594	< 2.22e-16	0.64	< 2.22e-16	0.852	< 2.22e-16	0.492	< 2.22e-16	0.591	< 2.22e-16	0.998
	0.525	0.558	< 2.22e-16	0.605	< 2.22e-16	0.644	< 2.22e-16	0.85	< 2.22e-16	0.497	< 2.22e-16	0.591	< 2.22e-16	0.998
	0.55	0.57	< 2.22e-16	0.601	< 2.22e-16	0.652	< 2.22e-16	0.855	< 2.22e-16	0.497	< 2.22e-16	0.597	< 2.22e-16	0.998
Cepstral coefficients	0.9	0.237	< 2.22e-16	0.731	< 2.22e-16	0.772	< 2.22e-16	0.912	< 2.22e-16	0.698	< 2.22e-16	0.727	< 2.22e-16	0.869
	0.925	0.221	< 2.22e-16	0.744	< 2.22e-16	0.783	< 2.22e-16	0.915	< 2.22e-16	0.694	< 2.22e-16	0.737	< 2.22e-16	0.869
	0.95	0.202	< 2.22e-16	0.771	< 2.22e-16	0.802	< 2.22e-16	0.916	< 2.22e-16	0.717	< 2.22e-16	0.758	< 2.22e-16	0.869
	0.975	0.161	< 2.22e-16	0.814	< 2.22e-16	0.831	< 2.22e-16	0.831	< 2.22e-16	0.733	< 2.22e-16	0.779	< 2.22e-16	0.869
SSS	0.025	0.459	< 2.22e-16	0.248	< 2.22e-16	0.745	< 2.22e-16	0.845	< 2.22e-16	0.615	< 2.22e-16	0.651	< 2.22e-16	0.998
	0.05	0.477	< 2.22e-16	0.194	< 2.22e-16	0.649	< 2.22e-16	0.827	< 2.22e-16	0.558	< 2.22e-16	0.614	< 2.22e-16	0.998
	0.075	0.517	< 2.22e-16	0.149	< 2.22e-16	0.615	< 2.22e-16	0.825	< 2.22e-16	0.529	< 2.22e-16	0.601	< 2.22e-16	0.998
	0.1	0.551	< 2.22e-16	0.092	7.0217e-06	0.591	< 2.22e-16	0.815	< 2.22e-16	0.506	< 2.22e-16	0.579	< 2.22e-16	0.998

Table 4. Results of the correlation tests between the different metrics and the IMSS on the mid-third of the vowel on artificial stimuli in terms of location, acoustic measurements and coverage.

In terms of coverage, most of the metrics can process almost every stimulus, i.e. more than 99% of the stimuli. However, the cepstral coefficients metric consistently shows lower coverage values, implying that only about 87% of the stimuli could be processed. Besides the IMSS based on cepstral coefficients, the coefficient of change also leads to lower coverage scores but only with the two lowest τ values.

As far as the temporal correlation coefficients are concerned, the starting point, center and end point determined by the metric-based IMSS are all significantly correlated to the starting point, center and end point based on the IMSS at the temporal center of the vowel. However, there is variation in the strength of the correlation, with the correlation coefficients ranging from 0.161 to 0.571 in terms of starting point, from 0.092 to 0.814 as to the center and from 0.413 to 0.831 as far as the end point is concerned. Most importantly, the two metrics which are less correlated to the traditional IMSS are the cepstral coefficients and the SSS in terms of starting point and center, respectively.

Turning to a detailed analysis of the acoustic measurement correlation coefficients, it can be observed that the different metrics are more correlated to the output of the mid-third IMSS in terms of F0 (between 0.709 and 0.916) than in terms of F1 and F2 (between 0.377 and 0.733 and between 0.448 and 0.779, respectively). Across F0, F1 and F2, the cepstral coefficient metric consistently leads to the highest correlation with the IMSS on the mid-third. The second highest correlation coefficients are obtained by the SSS, but generally speaking, the correlation coefficients are high for each metric and τ values.

6.3. Discussion

In experiment II, the procedures to identify spectral stability based on different metrics were compared with the most traditional IMSS technique, both in terms of the temporal location of the segment of the vowel identified as stable and in terms of subsequent acoustic measurements (F0, F1 and F2) carried out on those portions of the vowel. In addition, their coverage has also been calculated.

First of all, the most important question which interests phoneticians is whether using one technique or another would yield markedly different acoustic measurements compared to using the most traditional

IMSS on the center of the vowel. In response to this question, it is crucial to signal that no significant disparities arise. In other words, the F0, F1 and F2 measurements are all moderately to highly correlated (between 0.377 and 0.943) to the same measurements carried out on the center of the vowel. Although the slope in the log F2 – log F1 space yields slightly lower correlation coefficients than the other metrics, the correlation coefficients are rather similar across metrics for F0 (between 0.862 and 0.943), F1 (between 0.377 and 0.733) and F2 (between 0.448 and 0.779), meaning that choosing one or the other metric would not change to a large extent the subsequent measurements made on the portion identified as stable. This is especially true for F0. Most interestingly, the metrics that perform in Experiment I equally good as the traditional IMSS (i.e., cepstral coefficients) or even better (i.e., SSS) lead in most cases to the highest correlation coefficients among the different metrics.

Turning our attention to the location of the identified stable portions, it is worth signalling that the correlation coefficients of the temporal time points are lower than those of the acoustic measurements. Since the stable portions in those stimuli were randomly generated and given that the traditional IMSS on the center of the vowel is not flexible between stimuli, it is expected that if an IMSS metric accurately captures the stable portion of the vowel, which may not be centrally located, the correlation between the traditional IMSS and the other metric would be lower. The results show that all metrics correlate relatively with the traditional IMSS, indicating some overlap between all IMSS even if the location and duration of the vowel was made to vary randomly. Besides that, it is worth signalling that the two metrics that perform best in Experiment I (i.e., that are best at detecting the stable portion of the vowel), also lead to the lowest correlation coefficients, at least in terms of starting point for cepstral coefficients and center for SSS. In other words, it shows that those two metrics are the most different ones than the traditional IMSS in the way they capture the stable portion of vowels, most probably when the stable portion is situated further away from the temporal center of the vowel.

Regarding coverage, most techniques can easily cover a large majority of the stimuli, with over 99% of the stimuli handled by most of them. However, the cepstral coefficients metric only handles 87% of the stimuli, probably because it requires more frames for its computation than the other IMSS techniques. As to the lower coverage values obtained by the coefficient of change IMSS with two τ values, the reason for this remains unclear. Nevertheless, each metric can deal with a significant number of stimuli, at least 72%.

In conclusion, Experiment I demonstrates that the cepstral coefficients and SSS metrics perform as well as or even better than the traditional IMSS in detecting stable portions in the artificial stimuli. Nevertheless, the findings of Experiment II show that all metric-based IMSS are correlated to the traditional IMSS, indicating no large difference in the identification of the stable portion, even though the differences were larger for the metrics which performed best in Experiment I. Most importantly, the results also show that the choice of IMSS has very little impact on the subsequent acoustic measurements, indicating that the significant differences between metrics observed in Experiment I do not translate into practically large enough differences for phoneticians. 7. Experiment III - real speech data

Experiment III studies how the different metrics perform on real speech data. It addresses the question of whether the metrics affect the acoustic measurements which are carried out in the identified region of spectral stability.

7.1. Materials and methods

The speech data on which the metrics were tested were taken from a corpus of Belgian Dutch child speech. These data had been collected by means of a (non)word repetition task, which was first described in Verhoeven et al. (2016).

7.1.1. Stimuli and data selection

The database contains recordings of 36 monosyllabic (non)words which consisted of a vowel nucleus with one of the 12 monophthongs of Belgian Standard Dutch, i.e. [i, ʏ, ɪ, ɛ, ɑ, ɔ, u, yː, eː, øː, aː, oː] (Verhoeven, 2005). Each vowel occurred in three different consonantal contexts: (i) [p_t], (ii) [l_t] and (iii) [t_r]. Sixteen stimuli were existing Dutch words. The 20 other items were non-words which respect the requirements of the Dutch phonological system.

These (non)words had first been produced by a trained phonetician and native speaker of Standard Belgian Dutch. The recordings of these stimuli were played to the participating children who were asked to repeat the stimuli one by one. The recordings were made by means of a TASCAM DAT recorder and a head-mounted MicroMic II in a quiet room. The audio files were formatted to WAV-files by means of a TASCAM US 428 Digital Control Surface with a sampling rate of 44.1 kHz and a quantification precision of 16 bits per sample.

First, children's productions were perceptually assessed by six expert listeners to identify the vowels which were correct imitations of the target vowels. 7261 vowels out of 7985 were deemed correct imitations by the listening panel. In order to have a more between-children balanced corpus, a further selection was made by selecting children who produced at least 2 repetitions of the three cardinal vowels [i, aː, u] and at least one repetition of all vowels in whatever phonetic context. This amounted to a total of 4757 vocalic productions produced by 47 children.

7.1.2. Participants

The participating children were all Belgian Dutch-speaking born from native speakers of Belgian Standard Dutch. Their median chronological age was 6 years, with a minimum of 5 and a maximum of 7 years, and they all attended their first year of primary school. They had always lived in their region of birth before data collection. The normal-hearing status of the children was confirmed informally by reports from parents and teachers. No formal hearing test was carried out.

7.1.3. Acoustic analysis

All vowels were annotated by hand. The F0, F1 and F2 of all vowels were analysed by means of a Python script. The measurement of F0, F1 and F2 was carried out by the Parselmouth API (Jadoul et al., 2018) of PRAAT (Boersma and Weenink, 2021) via its standard auto-correlation algorithm. The pitch floor and ceiling were set to 150 and 500 Hz, respectively. The maximum number of candidates was set to 15, the silence threshold to 0.03, the voicing threshold to 0.45, the octave cost to 0.01, the octave-jump cost to 0.35, the voiced/unvoiced cost to 0.14. As far as the formant parameters are concerned, the maximum number of formants was set to 5, the maximum formant to 5500 Hz, the window length to 0.025 and the pre-emphasis to 50. The F0, F1 and F2 values of each vowel were measured as the means of all the measurements inside the selection determined by the different metrics.

7.1.4. Statistical analysis

When it comes to real speech data, two types of criteria can evaluate the output of the algorithms. On the one hand, it can be assessed to what extent the different algorithms correlate in terms of starting, middle and end points. On the other hand, it can be established whether the acoustic measurements carried out on the segmented audio signals differ from those obtained in the middle third portion of the vowels. To this end, F0, F1 and F2 in Hz were measured in each vowel. To compare the different outputs, Pearson's correlation coefficients were used with a significance level set at 0.05.

In addition to those two parameters, it is of interest to observe to what extent it is computationally possible to apply the different metrics on real speech stimuli, i.e. coverage.

7.2. Results

Table 5 surveys the correlation coefficients between the acoustic measurements carried out in the spectrally stable portion of the vowels as identified by the different metrics and the acoustic measurements carried out around the temporal center of the vowels. Table 5 also presents the correlation coefficients between the starting point, center and end point as identified by the different metrics and the starting point, center and end point of the IMSS on the mid-third. In addition, Table 5 shows the coverage for each selected metric and τ value.

Algorithm	τ	Starting point		Centre		End point		F0		F1		F2		Coverage
		r	*p*-value	r	*p*-value	r	*p*-value	r	*p*-value	r	*p*-value	r	*p*-value	
Coefficient of change	0.05	0.356	<2.22e-16	0.359	<2.22e-16	0.363	<2.22e-16	0.876	<2.22e-16	0.884	<2.22e-16	0.83	<2.22e-16	0.33
	0.075	0.314	<2.22e-16	0.321	<2.22e-16	0.327	<2.22e-16	0.879	<2.22e-16	0.901	<2.22e-16	0.839	<2.22e-16	0.55
	0.95	0.357	<2.22e-16	0.646	<2.22e-16	0.773	<2.22e-16	0.941	<2.22e-16	0.963	<2.22e-16	0.931	<2.22e-16	0.999
	0.975	0.232	<2.22e-16	0.696	<2.22e-16	0.837	<2.22e-16	0.952	<2.22e-16	0.973	<2.22e-16	0.942	<2.22e-16	0.999
Slope in log F2 – log F1 space	0.65	0.336	<2.22e-16	0.409	<2.22e-16	0.473	<2.22e-16	0.905	<2.22e-16	0.911	<2.22e-16	0.842	<2.22e-16	0.968
	0.675	0.311	<2.22e-16	0.388	<2.22e-16	0.457	<2.22e-16	0.904	<2.22e-16	0.914	<2.22e-16	0.843	<2.22e-16	0.971
	0.7	0.317	<2.22e-16	0.398	<2.22e-16	0.47	<2.22e-16	0.906	<2.22e-16	0.914	<2.22e-16	0.852	<2.22e-16	0.971
	0.725	0.314	<2.22e-16	0.4	<2.22e-16	0.475	<2.22e-16	0.907	<2.22e-16	0.919	<2.22e-16	0.854	<2.22e-16	0.971
Pooled within-variance	0.475	0.275	<2.22e-16	0.331	<2.22e-16	0.381	<2.22e-16	0.887	<2.22e-16	0.903	<2.22e-16	0.852	<2.22e-16	0.992
	0.5	0.257	<2.22e-16	0.316	<2.22e-16	0.369	<2.22e-16	0.889	<2.22e-16	0.907	<2.22e-16	0.853	<2.22e-16	0.995
	0.525	0.265	<2.22e-16	0.329	<2.22e-16	0.385	<2.22e-16	0.889	<2.22e-16	0.908	<2.22e-16	0.852	<2.22e-16	0.995
	0.55	0.246	<2.22e-16	0.313	<2.22e-16	0.373	<2.22e-16	0.892	<2.22e-16	0.91	<2.22e-16	0.855	<2.22e-16	0.995
Cepstral coefficients	0.9	0.308	<2.22e-16	0.471	<2.22e-16	0.557	<2.22e-16	0.944	<2.22e-16	0.978	<2.22e-16	0.956	<2.22e-16	0.523
	0.925	0.283	<2.22e-16	0.487	<2.22e-16	0.593	<2.22e-16	0.947	<2.22e-16	0.98	<2.22e-16	0.962	<2.22e-16	0.523
	0.95	0.223	<2.22e-16	0.504	<2.22e-16	0.638	<2.22e-16	0.953	<2.22e-16	0.981	<2.22e-16	0.969	<2.22e-16	0.523
	0.975	0.086	0.019766	0.542	<2.22e-16	0.723	<2.22e-16	0.957	<2.22e-16	0.981	<2.22e-16	0.969	<2.22e-16	0.523
SSS	0.025	0.274	<2.22e-16	0.519	<2.22e-16	0.575	<2.22e-16	0.94	<2.22e-16	0.968	<2.22e-16	0.942	<2.22e-16	0.999
	0.05	0.357	<2.22e-16	0.495	<2.22e-16	0.526	<2.22e-16	0.936	<2.22e-16	0.962	<2.22e-16	0.935	<2.22e-16	0.999
	0.075	0.381	<2.22e-16	0.481	<2.22e-16	0.505	<2.22e-16	0.931	<2.22e-16	0.957	<2.22e-16	0.926	<2.22e-16	0.999
	0.1	0.3999	<2.22e-16	0.472	<2.22e-16	0.489	<2.22e-16	0.93	<2.22e-16	0.954	<2.22e-16	0.921	<2.22e-16	0.999

Table 5. Results of the correlation tests between the different metrics and the IMSS on the mid-third of the vowel on real speech data in terms of location, acoustic measurements and coverage.

As far as coverage is concerned, the metric which makes it possible to process the largest number of stimuli is the SSS, while the cepstral coefficient metric consistently triggers one of the lowest coverages among the tested metrics. The pooled within-variance and the slope in log F2 - log F1 space shows excellent coverage, i.e. they can process more than 95% of the stimuli. The coefficient of change exhibits a relatively small coverage with the smaller τ values and good coverage with the higher τ values.

It can also be seen that the correlation coefficients of the temporal time points are lower than those of the acoustic measurements. In other words, the effect of the different metrics is observable at the level of the location of the spectrally stable portion. Fig. 11 shows the distribution of the relative location of the center of the spectrally stable portion relative to the total duration of the vowel. As a first observation, the different τ thresholds do not have a large influence on the shape of the distribution per metric but it has an effect on the number of observations.

The metrics used by Hillenbrand et al. (1995) and Van Bergem (1988) show a clear bi-modal distribution with the two peaks being situated towards the edge of the vowel. This indicates that these metrics tend to locate the spectrally stable part of the vowel towards the beginning or the end of the vowel. The distribution based on the coefficient of change (Lennig, 1978) also exhibits a bi-modal distribution whose peaks are more centerd. The metric using cepstral coefficients shows a bi-modal distribution, but most of the observations are closer to the center of the vowel. Finally, the SSS shows a right-skewed distribution.

Turning to a detailed analysis of the correlation coefficients, it can be observed that the different metrics are more correlated to the output of the mid-third IMSS in terms of center and end point (between 0.316 and 0.696 and between 0.327 and 0.837, respectively) than in terms of starting point (between 0.086 and 0.399). The higher correlation in terms of starting point results from using the SSS metric, while the highest correlations in terms of the center and end point are obtained by the coefficient of change metric.

Furthermore, the acoustic measurements are more strongly correlated irrespective of the metric. This means that feature-defined IMSS can differ to some extent from the traditional IMSS around the middle of the vowel, but it also means that the different IMSS procedures do not have such an effect on the subsequent acoustic measurements made on the identified stable portion. As far as the acoustic measurements are concerned, it can be observed that all metrics have very high correlation coefficients. For

F0, F1 and F2, the metrics which consistently across τ values lead to one of the highest correlation coefficients are the SSS and the cepstral coefficients.

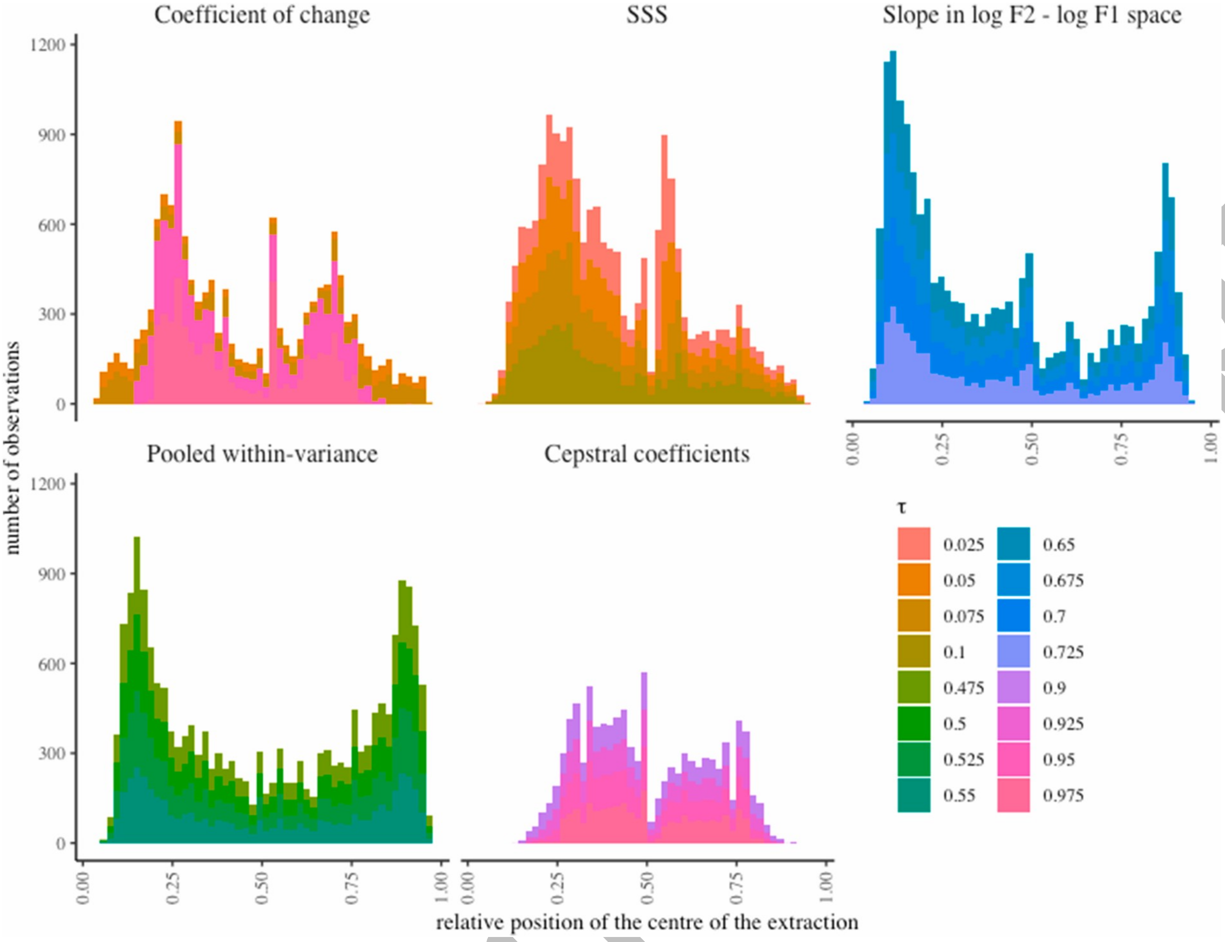


Fig. 11. Distribution of the relative position of the center of the IMSS as a function of metrics and τ value.

7.3. Discussion

In this experiment, the procedures to identify spectral stability based on different metrics were tested on natural speech data. First, F0, F1 and F2 were measured in the portions of the vowels identified as spectrally stable by different algorithms; these measurements were compared to the measurements made in the middle of the vowel. The distribution of the locations of the center of the spectrally stable portions was also analysed in order to observe differences in the location of the spectrally stable portions according to the different metrics. The coverage on real speech data has also been calculated.

The first question to be addressed relates to the potential effect of using an IMSS procedure based on a given metric rather than a more traditional method for selecting the stable part of the vowel. In other words, does the use of a specific metric for the identification of the stable portion have an impact on the subsequent acoustic measurements? From Table 5, it is clear that the correlation between measurements in all algorithms and those in the middle of the vowel is very high. This suggests that all methods provide measurements which genuinely represent the actual vowel target.

Interestingly enough, the two feature-defined metrics which were best at detecting the stable portion of the vowel in synthesized stimuli correlate better with the IMSS on the mid-third than with the other metrics. That is, the results obtained by metrics which are known to provide the most reliable results (according to Experiment I) are very similar to those obtained by the IMSS which selects the mid-third of the vowel. In other words, by using the mid-third IMSS method, F0, F1 and F2 measurements are expected to be very close to the measurements which would be made on the spectrally stable part of the vowel if it were to be known.

Turning to the location of the center of the selection, the IMSS based on the SSS shows a right-skewed distribution. This means that this metric, which performs best at detecting the spectrally stable portion of a vowel, locates the center of the spectrally stable portion in the first quarter or third of the vowel. This finding is similar to Evanini (2009). Nevertheless, a certain number of spectrally stable sections are located sparsely through the rest of the vowel, especially in the vicinity of its center. The bi-modal distribution of the centers obtained by the IMSS based on the cepstral coefficients also seems to show that more observations are situated within the first half of the vowel than in the second half. Nevertheless, it can be seen that most of the observations are situated towards the middle of the vowel rather than towards the edges. On the contrary, the bi-modal distributions of the three other metrics that performed worse on synthesized stimuli seem to identify the center of the spectrally stable portion towards the edges. A possible explanation for this might be that those metrics can be sensitive to the presence of outliers in terms of frame instability around the center of the vowel. Another possible explanation might be that, during a potential long voiced occlusion, the metrics which all use formant values capture some stability in the formants value during the consonant-vowel transitions. The amount of stability would then be superior to the one observed within the middle of the vowel. Therefore, those algorithms might erroneously prefer to select portions of the signal situated before or after the stable portion of the vowels.

The metric which can be applied to the largest number of stimuli is the SSS. This simple algorithm performs with good accuracy and can be applied to almost every vowel. Other metrics seem to perform quite well in terms of coverage too. Most of them can cover more than 95% of the stimuli, while the use of the cepstral coefficients can only handle 52% of the stimuli in this specific test. A possible explanation for this is that the ease of computation of the LTAS hardly ever prevents the IMSS, whereas the cepstral coefficient metric suggested by Van Bergem (1993) requires five frames to be used. It should be noted at this point that the good correlation achieved by the MFCC-based IMSS procedures are obtained on the stimuli which were processable, i.e. longer stimuli or stimuli with a more easily identifiable spectrally stable portion. Therefore, audio samples whose duration is too short cannot be processed by those algorithms. The relatively smaller coverage of the formant-based IMSS procedures might be the result of the inability of formant tracking to be carried out on some audio samples. More specifically, the coefficient of change with a low τ value also seems to perform worse. It is probably too restrictive in terms of variability tolerated and thus cannot select a sequence of at least two frames whose instability is below τ.

8. General discussion

In this research, even though one has to cope with the limitations of both synthesized and natural speech data, the adopted methodology makes it possible to observe in as controlled a setting as possible, both the accuracy of different IMSS techniques and their effect on subsequent acoustic measurements.

In Experiment I, it has been shown that the metrics which prove to be better at detecting the stable portion of a vowel are those which use higher dimensional representations of spectra, such as the cepstral coefficients or the SSS. Some metrics manage to get logits of f1-scores equal to the performance of the traditional IMSS procedure around the center of the vowel. The cepstral coefficients metric performs significantly better with the most extreme τ value tested, i.e. $\tau = 0.975$. Similarly, the SSS performs significantly better with the four tested τ values. On the contrary, the formant-based metrics do not reach the level of performance of the IMSS on the mid-third, they are even worse in most cases.

Experiment II has shown that, although highly significant differences can be observed in the accuracy of the IMSS (as in Experiment I), the subsequent acoustic measurements are very similar, whatever metrics is used. Of particular significance to phoneticians is the finding from Experiment III, which demonstrates that when applying these IMSS techniques to real speech, the differences between them in terms of acoustic measures are even further reduced. This suggests that no substantial advantage of specific metrics over the traditional method was observed.

Fig. 12 indicates that the algorithms which perform best (i.e., higher estimates) are associated with higher correlation coefficients. In other words, algorithms which are good at detecting the spectrally stable vowel portions provide acoustic measurements closer to those obtained by the traditional IMSS around the center of the vowel. It can also be seen that when the τ of each algorithm improves the IMSS on artificial stimuli, higher correlation is achieved with the measurements obtained via the traditional IMSS. That trend is not observable with the coefficient of change. The reason for this is still unclear. Generally speaking, it means that optimisation of the algorithm is likely to provide better correlations. All in all, it indicates that improving the IMSS will reflect in lower impact on the acoustic measurements.

Besides the more or less equivalent performance of the metrics, its effect on the F0, F1 and F2 measures is limited. It means that, despite using a flexible algorithm which can capture to some extent the stable portion of the vowel, the use of a very simple algorithm which selects the middle portion of the vowel produces very similar results. To put it simply, using metrics such as the SSS or cepstral coefficients might improve the identification of the spectrally most stable portion of the vowel compared to the traditional IMSS. As a result, this leads to slightly different temporal measures observed on artificial stimuli. However, the differences in terms of acoustic measurements (e.g. F0, F1 and F2) are practically small, especially on real speech data. In other words, the best fine-tuned IMSS techniques might result in a slight improvement of the IMSS, but they would not lead to significantly relevant differences in terms of acoustic measurements, with the most traditional method leading essentially to the same results.

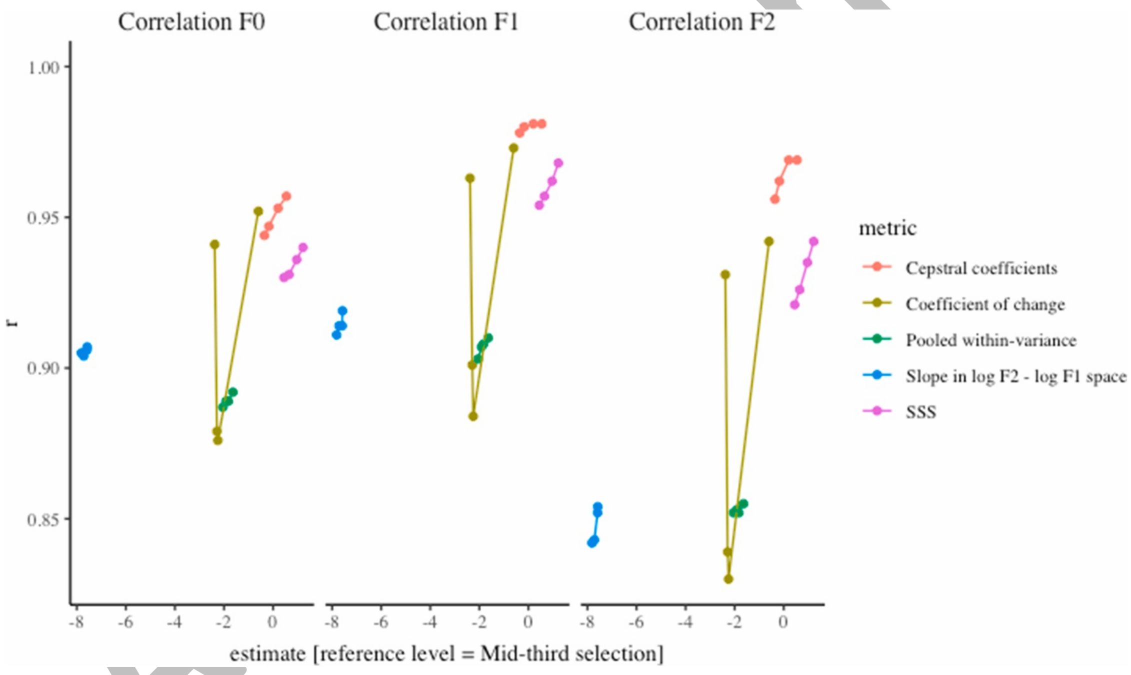


Fig. 12. Correlation coefficients for F0, F1 and F2 measured on real speech data as a function of the estimate obtained on synthetic stimuli according to the different metrics.

8.1. Limitations and future directions

This research shows the efficiency of different metrics to capture spectral stability and the effect of this on subsequent acoustic measurements. It may be useful, however, to point out some limitations due to the methodological choices which were made.

Firstly, the evaluation carried out on synthetic stimuli does not indicate whether the algorithm performs the same way with real speech data, but the nature of real speech data prevents having a gold standard to evaluate the performance of the metrics. Secondly, the algorithms were tested on non-spontaneous speech collected in carefully controlled conditions. The application of the algorithm to spontaneous speech with a potentially less controlled paradigm has not been tested. Thirdly, to compare the different metrics the same window duration (20 ms) and rate (5 ms) were used in all cases. It is not clear whether those settings might suit some algorithms better than others. For instance, the metric suggested by Van Bergem (1993) requires a larger number of frames. A very short vowel may thus be difficult to process. Shorter windows might be more suitable for such metrics. For each metric, an ideal trade-off between the time resolution and the metric resolution should be found. The adjustment of such parameters is beyond the scope of the present study, but further research in this area seems useful. Moreover, only rectangular windows were used in this first study. Some other window types might further improve the IMSS method.

The present paper focuses on the metrics themselves, but *a posteriori* treatments of the metrics might further improve the ability of an algorithm to detect the spectrally stable portion of speech sounds. Therefore, further research on the use of probabilistic and deep-learning approaches from speech processing to detect the stable portion of vowels is suggested.

Furthermore, this research is based on the assumption that only one measurement is used to characterize a vowel. However, rather than collecting one data point per vowel, it would be possible to collect multiple data points in the form of a time-series and to analyse the vowel as a whole by using, for instance, General Additive Mixed Modelling (see Wieling, 2018). The ability of GAMMs to deal with dynamic data makes it possible to separate between parametric terms and random smooths in order to separate a height effect from its (non)linear pattern. Yet, the reference value for the fixed effect of time on the acoustic measurements needs to be centered in order for it to be representative of the vowel. In light of the above, it can reasonably be assumed that centring the effect of time around the center of the vowel gives results representative of the vowel target.

9. Conclusions

This study compared and evaluated different metrics that have been used previously in research to evaluate spectral stability and their advantages and disadvantages were discussed in a theoretical perspective. A novel metric has also been developed to improve the identification of spectral stability. The reliability of the different metrics has been investigated experimentally.

First, the ability of the metrics to capture spectral stability in synthetic stimuli was evaluated. All metrics were assessed with respect to synthesized speech stimuli whose spectrally stable part had been defined beforehand. This experiment showed that feature-defined metrics hardly reach the performance of the technique which extracts the middle portion of a vowel. The results also suggest that the larger the frequency range covered by the metric, the better it captures spectral stability. Secondly, the effect of the metrics on acoustic measurements was examined on real speech data. The results show that a metric which performs better gets results which are very similar to the IMSS of the mid-third.

In short, this research provides empirical validation for accepting the middle of the vowel as stable is a good trade-off between ease of implementation and reliability. Other metrics such as cepstral coefficients or the SSS can be used in order for the IMSS to be more flexible but overall, they hardly achieve better results than the traditional approach. Those fine-grained IMSS techniques can differ in the portions of the vowel that they detect as stable, but the subsequent acoustic measurements made on the identified stable portions are anyway very similar to each other, whatever metric is used. Furthermore, the feature- defined algorithms which perform best on synthetic stimuli yield acoustic measurements which are the most highly correlated to the ones obtained the middle of the vowel. In this respect, it is reassuring that a widely-used technique provides reliable results by selecting portions of the acoustic signal which represent the vowels well.

CRediT authorship contribution statement

Jérémy Genette: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft, Visualization. Jose Manuel Rivera Espejo: Formal analysis. Steven Gillis: Conceptualization, Methodology, Formal analysis, Data curation, Writing – original draft. Jo Verhoeven: Conceptualization, Methodology, Formal analysis, Data curation, Writing – original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

Thanks are due to the participants of this study. The work of the second author was funded by the Special Research Fund (BOF) of the University of Antwerp [ID 44208]. The work of the other authors was supported by the Research Foundation in Flanders (FWO) [grant G004321N].

Appendix A. Formulae

Please note that the present implementation differs to some extent from the original applications of the metrics because the parameters were (manually) set by the original authors to optimize the output of a given metric, but this would not provide a reliable basis for comparison. In fact, setting the parameters to improve the accuracy of a metric gives information about the level of accuracy that can be reached with that metric on a given test set, but it gives little objective information about the ability to evaluate spectral stability by the metric *per se*.

1. Lennig (1978)

It computes an instability score, that is, the coefficient of change presented in Eq. (1) where c_f is the coefficient of change of frame *f*.

$$c_f = \frac{|F1_f + F1_{f-1}| + |F1_f + F1_{f+1}|}{F1_f} + \frac{|F2_f + F2_{f-1}| + |F2_f + F2_{f+1}|}{F2_f} \qquad (1)$$

2. Hillenbrand et al. (1995)

It consists in selecting the frames with the lowest slope in the log F2 - log F1 space as in Eq. (2) where $slope_f$ is the slope in log F2 - log F1 space of frame *f* and where *sf* stands for the start of frame *f* and *ef*, the end of frame *f*.

$$slope_f = \frac{\log F2_{ef} - \log F2_{sf}}{\log F1_{ef} - \log F1_{sf}} \tag{2}$$

3. Van Bergem (1988)

$$pv_f = \frac{(n-1) * \frac{\sum_{i=1}^{n}\left(\log F1_f - \overline{x}\right)^2}{n} + (m-1) * \frac{\sum_{i=1}^{m}\left(\log\ F2_f - \overline{x}\right)^2}{m} + (p-1) * \frac{\sum_{i=1}^{p}\left(\log\ F3_f - \overline{x}\right)^2}{p}}{n + m + p\ -\ 3} \tag{3}$$

It computes an instability score as in Eq. (3) where pv_f is the pooled within-variance of the log-transformed formant values within frame *f*.

4. Van Bergem (1993)

It consists in the within standard deviation for the frame interval whose center is frame *f* as in Eq. (4) where N stands for the number of cepstral coefficients calculated within a frame interval, c_{ij} represents the j^{th} cepstral coefficient of frame *i* and $\overline{c_j}$ is the mean value of all the j^{th} cepstral coefficients of the frame interval.

$$V(f)\ =\ \sqrt{\frac{1}{N}\sum_{i=f-5}^{f+5}\sum_{j=1}^{8}(c_{ij} - \overline{c_j})} \tag{4}$$

Appendix B. Articulatory targets specification

The articulatory target specifications are adapted from Boersma (1998) and the PRAAT scripts available on its companion website. To synthesize the CVC sequences, a contraction of the lungs is synthesized by setting the *Lung* parameter to 0.2 at 0 ms and to 0.0 after 100 ms in order to generate a sufficient lung pressure which triggers an increase in oral pressure necessary for phonation and the production of stops (see Boersma, 1998: 128). This parameter is the same for all stimuli.

To generate an [a]-like sound, vocal fold vibration is generated by setting the *Interarytenoid* parameter to 0.5. To avoid nasalization, the activity of the *LevatorPalatini* is set at 1.0 in order to close the nasal port. The pulling-down of the tongue is synthesized by means of *Hyoglossus* activity of 0.5. and jaw opening by a *Masseter* activity of -0.5.

In order to synthesize [b]-like sounds, the *Interarytenoid* target is set to 0.53, i.e. 0.5 for voicing and an additional activity of 0.03 is added in order to compensate for the oral closure. The *LevatorPalatini* is also set to 1 in order to prevent air leakage through the nasal cavity. The *Masseter* parameter is set to 0.5 to lift the jaw. The *OrbicularisOris* activity is set to 1 to synthesize lip closure. The *Hyoglossus, UpperTongue* and *Styloglossus* are kept in their neutral position, i.e. 0. The synthesis of a [p]-like sound requires the same articulatory targets, except for the *Interarytenoid* target, which is set to 0. To synthesize an [m]-like sound, the same parameters as for the [b] are used except for the *LevatorPalatini* which is set to 0 in order to synthesize a lowering of the velum.

For the synthesis of [d]-like sounds, the *Interarytenoid* target is set to 0.53 and the *LevatorPalatini* parameter to 1 for the same reasons as those evoked with respect to the synthesis of [b]-like sounds. However, the *OrbicularisOris* and *Masseter* activity are set to 0. The articulator used to synthesize the oral closure in [d]-like sounds is the *UpperTongue* whose activity is set to 1. To synthesize a [t]-like sound, the same parameters are used, but the *Interarytenoid* parameter is set to 0. As far as [n]-like sounds are concerned, the synthesis requires the same parameters as for the [d]-like sounds, but the lowering of the velum is synthesized by setting the *LevatorPalatini* activity to 0.

[g]-like sounds are also synthesized. The *Interarytenoid* target is thus set to 0.53 and the *LevatorPalatini* parameter to 1. The *OrbicularisOris* and *Masseter* activity are set to 0. The velar closure is realized by setting the *Styloglossus* activity to 1. The *Hyoglossus* and *UpperTongue* activity are set to 0. To synthesize [k]-like sounds, the same parameters are used, but the *Interarytenoid* activity is set to 0. The synthesis of [ŋ]-like sounds uses the same parameters as the synthesis of [g]-like sounds but with an activity of the *LevatorPalatini* reduced to 0.

Appendix C. Timing targets specification

The sequences are built in such a way that there is one point (OccC1) where all the articulators needed for the articulation of C1 have reached their target and did not depart from it. Another time point is located where the articulator has reached its target (TsOccC1) before OccC1. After OccC1, there is another point (TeOccC1) where the level of activity of one articulator that characterizes the C1 target starts varying progressively towards the level of activity that characterizes the articulation of the vowel. The same principle applies to all articulators. The vocalic target is reached at TsV and is maintained up to TeV. The rules for the synthesis of the C1-V articulation apply symmetrically to the V-C2 articulation. Between two targets, a varying in-between position of the articulator is synthesized which we assume to be somehow similar to transition movements involved in coarticulation.

In practice, the first time point which is determined is the onset of the vocalic stable portion (TsV). It needs to be situated between 50 ms after the beginning of the sound file and at least 100 ms before its end. The end of the stable portion of the vowel (TeV) has to occur at least 50 ms after the beginning of the stable portion. Another time point is set between the onset of the recording and the start of the stable portion of the vowel to indicate the point at which all the articulators needed for the articulation of the C1 have reached their target (OccC1). A similar time point is set between the end of the stable portion of the vowel and the end of the sound to indicate that the point at which all the articulators needed for the articulation of C2 have reached their articulatory targets (OccC2). A time point is created between the beginning of the sound and the moment at which an articulator start moving towards its C1 target (Onset). Up to that point, the articulator keeps its neutral position. A similar time point (Offset) is situated between the moment at which an articulator leaves its C2 target and the end of the sound. By this time, the articulator has reached its neutral position.

If a given articulator has the same target in C1, V and C2, its onset point is set between the onset of the recording and C1 occlusion. It means that it must have reached its target before the occlusion of C1. Its offset point is situated between the occlusion of C2 and the end of the recording. It departs from its C2 target after C2 occlusion.

If an articulator has different C1, V targets and different V and C2 targets, its onset point is set between the start of the recording and the C1 occlusion, it reaches its target before C1 occlusion, where another time point is set, then moves away from it after C1 occlusion, where another time point is set, and before the start of the onset of the stable portion of V by which it has reached its target for the articulation of V. It departs from it at the end of the stable portion of V and reaches the target of C2 before C2 occlusion, where another time point is set. It then departs from it between the occlusion of C2 and the end of the recording, and eventually reaches its neutral position between C2 occlusion and the end of the recording, where another time point is set.

If an articulator has the same C1 and V targets but a different C2 target. Its onset point is set between the onset of the recording and the C1 occlusion, it reaches its target before C1 occlusion, then moves away from it at the end of the stable portion of V and before C2 occlusion. At this point it reaches its C2 target. The articulator departs from it after C2 occlusion and reaches its offset point between the departure from its C2 target and the end of the sound.

If an articulator has the same C2 and V targets but a different C1 target, its onset point is set between the onset of the recording and C1 occlusion, it reaches its target before C1 occlusion, then moves away from it after C1 occlusion and before the start of the stable portion of the vowel by which it has reached the target for the articulation of V. The articulator then departs from it after C2 occlusion and reaches its offset between that its departure from C2 occlusion and the end of the sound.

- Start: always 0 s;
- Onset: between Start and OccC1;
- TsOccC1: between Onset and OccC1;
- OccC1: between Start and TsV;
- TeOccC1: between OccC1 and TsV;
- TsV: after 0 + 0.05 s and before 0.35s-0.1 s;
- TeV: after TsV+0.05 s and before 0.35s-0.05 s;
- TsOccC2: between TeV and OccC2;
- OccC2: between TeV and End;
- TeOccC2: between OccC2 and Offset;
- Offset: between OccC2 and End;
- End: always 0.35 s.

References

Bates, D., Machler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models ¨ using lme4. J. Stat. Softw. 67 https://doi.org/10.18637/jss.v067.i01.

Boersma, P., 1998. Functional Phonology: Formalizing the Interactions between Articulatory and Perceptual Drives [PhD Thesis]. University of Amsterdam.

Boersma, P., Weenink, D., 2021. Praat: doing phonetics by computer [computer program. Version 6].

Buckland, M., Gey, F., 1994. The relationship between recall and precision. J. Am. Soc. Inf. Sci. 45, 12–19. https://doi.org/10.1002/(SICI)1097-4571(199401)45:1<12:: AID-ASI2>3.0.CO;2-L.

Derdemezis, E., Vorperian, H.K., Kent, R.D., Fourakis, M., Reinicke, E.L., Bolt, D.M., 2016. Optimizing vowel formant measurements in four acoustic analysis systems for diverse speaker groups. Am. J. Speech Lang. Pathol. 25, 335–354. https://doi.org/ 10.1044/2015_AJSLP-15-0020.

Development Core Team, 2021. R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria [Online] Available. https:// www.R-project.org/.

Duckworth, M., McDougall, K., Jong, G.de, Shockey, L., 2011. Improving the consistency of formant measurement. Int. J. Speech Lang. Law 18, 35–51. https://doi.org/ 10.1558/ijsll.v18i1.35.

Eichhorn, J.T., Kent, R.D., Austin, D., Vorperian, H.K., 2018. Effects of aging on vocal fundamental frequency and vowel formants in men and women. J. Voice 32, 644. e1–644.e9. https://doi.org/10.1016/j.jvoice.2017.08.003.

Embarki, M., Dodane, C., 2011. La coarticulation: Des indices a la Représentation. L'Harmattan.

Evanini, K., 2009. The Permeability of Dialect Boundaries [PhD Thesis]. University of Pennsylvania.

Farnetani, E., Recasens, D., Hardcastle, W.J., Laver, J., Gibbon, F.E., 2010. Coarticulation and connected speech processes. The Handbook of Phonetic Sciences. John Wiley & Sons, Ltd, pp. 316–352. https://doi.org/10.1002/9781444317251.ch9.

Fletcher, A.R., McAuliffe, M.J., Lansford, K.L., Liss, J.M., 2015. The relationship between speech segment duration and vowel centralization in a group of older speakers. J. Acoust. Soc. Am. 138, 2132–2139. https://doi.org/10.1121/1.4930563.

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. https://doi.org/10.1044/2016_JSLHR-S-15-0355.

Hillenbrand, J., Getty, L., Clark, M., Wheeler, K., 1995. Acoustic characteristics of American English vowels. J. Acoust. Soc. Am. 97, 3099–3111. https://doi.org/ 10.1121/1.409456.

Jadoul, Y., Thompson, B., Boer, B., 2018. Introducing Parselmouth: a python interface to Praat. J. Phon. 71, 1–15. https://doi.org/10.1016/j.wocn.2018.07.001.

Kent, R.D., Vorperian, H.K., 2018. Static measurements of vowel formant frequencies and bandwidths: a review. J. Commun. Disord. 74, 74–97. https://doi.org/10.1016/j. jcomdis.2018.05.004.

Kühnert, B., Nolan, F., Hewlett, N., Hardcastle, W.J., 1999. The origin of coarticulation.

Coarticulation: Theory. Data and Techniques, pp. 7–30. https://doi.org/10.1017/ CBO9780511486395.002.

Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B., 2015. Package ‘lmertest. R Package Version 2, 734.

Labov, W., Ash, S., Boberg, C., 2006. The Atlas of North American English: Phonetics, Phonology, and Sound Change: a multimedia reference tool. Mouton de Gruyter.

Lennig, M., 1978. Acoustic Measurement of Linguistic Change: The Modern Paris Vowel System [PhD Thesis]. University of Pennsylvania.

Lindblom, B.E., Studdert-Kennedy, M., 1967. On the role of formant transitions in vowel recognition. J. Acoust. Soc. Am. 42, 830–843. https://doi.org/10.1121/1.1910655.

Mennen, I., Scobbie, J.M., de Leeuw, E., Schaeffler, S., Schaeffler, F., 2010. Measuring language-specific phonetic settings. Second Lang. Res. 26 (1), 13–41. https://doi. org/10.1177/0267658309337617.

Miller, J.D., 1989. Auditory-perceptual interpretation of the vowel. J. Acoust. Soc. Am. 85, 2114–2134. https://doi.org/10.1121/1.397862.

Nadeu, C., Macho, D., Hernando, J., 2001. Time and frequency filtering of filter-bank energies for robust HMM speech recognition. Speech Commun. 34, 93–114. https:// doi.org/10.1016/S0167-6393(00)00048-0.

Van Bergem, D.R., 1988. The first step to a better understanding of vowel reduction. In: Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam, pp. 61–75.

Van Bergem, D.R., 1993. Acoustic vowel reduction as a function of sentence accent, word stress, and word class. Speech Commun. 12, 1–23. https://doi.org/10.1016/0167- 6393(93)90015-D.

Van der Harst, S., 2011. The Vowel Space Paradox: A Sociophonetic Study on Dutch. LOT.

Verhoeven, J., 2005. Belgian standard Dutch. J. Int. Phon. Assoc. 35, 243–247. https:// doi.org/10.1017/S0025100305002173.

Verhoeven, J., Hide, O., Maeyer, S., Gillis, S., Gillis, S., 2016. Hearing impairment and vowel production. A comparison between normally hearing, hearing-aided and cochlear implanted Dutch children. J. Commun. Disord. 59, 24–39. https://doi.org/ 10.1016/j.jcomdis.2015.10.007.

Weismer, G., Berry, J., 2003. Effects of speaking rate on second formant trajectories of selected vocalic nuclei. J. Acoust. Soc. Am. 113, 3362–3378. https://doi.org/ 10.1121/1.1572142.

Wieling, M., 2018. Analyzing dynamic phonetic data using generalized additive mixed modeling: a tutorial focusing on articulatory differences between L1 and L2 speakers of English. J. Phon. 70, 86–116. https://doi.org/10.1016/j.wocn.2018.03.002.