Systematic comparison of BIC-based speaker segmentation systems

Vassiliki Moschou, Margarita Kotti, Emmanouil Benetos, and Constantine Kotropoulos
Department of Informatics, Aristotle Univ. of Thessaloniki
Box 451, Thessaloniki 541 24, Greece
{vmoshou,mkotti,empeneto,costas}@aiia.csd.auth.gr

***Abstract*— Unsupervised speaker change detection is addressed in this paper. Three speaker segmentation systems are examined. The first system investigates the AudioSpectrumCentroid and the AudioWaveformEnvelope features, implements a dynamic fusion scheme, and applies the Bayesian Information Criterion (BIC). The second system consists of three modules. In the first module, a second-order statistic-measure is extracted; the Euclidean distance and the T^2 Hotelling statistic are applied sequentially in the second module; and BIC is utilized in the third module. The third system, first uses a metric-based approach, in order to detect potential speaker change points, and then the BIC criterion is applied to validate the previously detected change points. Experiments are carried out on a dataset, which is created by concatenating speakers from the TIMIT database. A systematic performance comparison among the three systems is carried out by means of one-way ANOVA method and post hoc Tukey's method.**

I. Introduction

Automatic speech segmentation aims at finding the speaker change points in an audio stream. Bayesian Information Criterion (BIC) has been widely applied for speaker segmentation [1], [2], [3], [4]. A two-pass segmentation technique called DISTBIC has improved the performance by utilizing distance-based segmentation before applying BIC [1]. A BIC alternative, which does not need tuning was introduced in [2]. Meanwhile, novel features like the smoothed zero crossing rate (SZCR), the perceptual minimum variance distortionless response, and the filterbank log coefficients were proposed [5]. A hybrid algorithm, which combines metric-based segmentation with BIC and model-based segmentation with Hidden Markov Models (HMMs) is described in [3]. A creditable review on speaker segmentation can be found in [6].

The major contribution of this paper is in the systematic comparison of the performance of three speaker segmentation systems performance. All systems are BIC-based, their efficiency is tested on the same dataset, and the same experimental protocol is utilized. The novelty of the paper lies in the fact that their performance is statistically analyzed by using one-way analysis of variance (one-way ANOVA), followed by Tukey's method. The first system investigates scalar and vector features and applies a fusion scheme, which combines the partial results so as it boosts efficiency. In the second system there are three modules. The first module preprocesses the utterances, the second module uses the Euclidean distance followed by the T^2 Hotelling statistic on MFCCs, and the last module utilizes BIC in conjunction with MFCCs. In the third system, the processing is split into two main stages: In the first stage, a metric-based approach using Line Spectral Pairs (LSPs) is implemented, while in the second stage, the detected potential speaker change points are validated using BIC.

The rest of the paper is organized as follows. The three systems are described in Section II. Experimental results are shown in Section III. Statistical analysis is detailed in Section IV, and, finally, conclusions are drawn in Section V.

II. BIC-based speaker segmentation

A. The first system

The system relies on the BIC variant proposed in [2]. The following features are extracted: the MFCCs; the maximum magnitude of the DFT coefficients in a speech frame; the short-time energy (STE); the AudioSpectrumCentroid; and AudioWaveformEnvelope [4]. The last two features belong to low level MPEG-7 descriptors. With respect to the representation of the speakers, every speaker is described with a multivariate Gaussian probability density function (pdf) with mean vector $\mathbf{m}$ and covariance matrix $\mathbf{\Sigma}$.

Multiple passes are allowed. In the first four passes, we resort to MFCCs; in the 5th pass the maximum DFT magnitude is used; in the 6th pass the STE is taken into account; in the 7th pass the MFCCs are explored again; in the 8th pass the AudioSpectrumCentroid is used; in the 9th pass the maximum DFT magnitude is investigated, and in the last pass the maximum of the AudioWaveformEnvelope is exploited. The decisions taken in one pass are fed to the next pass, as in a Bayesian network. After each pass, the number of chunks is decreased, due to specific potential change points are discarded as being false. Several researchers [1], [5], [11] have come to the conclusion that the more data are available, the better is the performance.

Referring to the scalar features, that is to the maximum magnitude of DFT, the STE, and the maximum of AudioWaveformEnvelope, a dynamic thresholding is employed. We start with an ad hoc threshold ϑ. It is determined after a considerable number of experiments so as to maximize the F_1 measure, as it is defined in (8). Let us consider a recording that has *I* chunks. We test the possible speaker change point c_j which lays between chunks k and $k+1$. If $f(k)$ is the current feature value computed at chunk k, we estimate $f(k+1)$ and then we calculate the value of the absolute difference

between these values denoted by $\epsilon = |f(k+1) - f(k)|$. Let $\bar{\epsilon}$ be the mean value of ϵ over all chunks of a recording: $\bar{\epsilon} = \frac{1}{I-1}\sum_{l=1}^{I-1}|f(l+1) - f(l)|$. Then $\bar{\epsilon}$ is compared to ϑ, whose value is adjusted as follows:

$$\vartheta' = \begin{cases} \vartheta + 0.005\bar{\epsilon} & \text{when } \vartheta < \bar{\epsilon} \\ \vartheta - 0.005\bar{\epsilon} & \text{otherwise.} \end{cases} \quad (1)$$

Whenever a feature vector is employed (such as the MFCCs), BIC is used. To estimate the GMM within BIC, the expectation-maximization (EM) algorithm is applied. However, the EM algorithm may converge at local minima. This issue, combined with the fact that BIC is a weak classifier, leads us to propose a fusion scheme. Thus, we could theoretically reduce the error by repeating the experiment R times and applying majority voting. To be more specific, we obtain a set of possible speaker turn points for each repetition. The possible speaker change points that make their appearance at a sufficient frequency S in the current pass are candidate turn points for the next pass. If they fail to appear S times, they are discarded. Both R and S are determined heuristically. Typical values for R and S are 5 and 4, respectively. The aforementioned procedure is detailed in [4].

B. The second system

The second system is structured in three modules. In the first module, a total set of 24 features per segment is extracted. The set includes the mean and the variance of the following feature values: maximum magnitude of DFT, STE per segment, maximum of AudioWaveformEnvelope per segment, and maximum of AudioSpectrumCentroid per segment. It also considers their first-order (delta) and second-order (delta-delta) differences. A feature selection algorithm is applied in order to derive the optimum feature subset for speaker segmentation [7]. The search strategy implemented is Branch and Bound, which has an almost optimal performance. The traversing algorithm uses depth first search with a backtracking mechanism. The criterion utilized for selection is $J = \mathrm{tr}(\mathbf{S}_w^{-1}\mathbf{S}_b)$ where $\mathrm{tr}(\cdot)$ stands for the matrix trace operator, $\boldsymbol{S}_w$ is the within-class scatter matrix, $\boldsymbol{S}_b$ is the between-class scatter matrix. The best 5 out of the 24 features are selected. Starting from the most efficient, the selected features are: the mean magnitude of the DFT, the delta AudioWaveformEnvelope, the mean STE, the AudioWaveformEnvelope, and the variance of the delta magnitude of the DFT.

Next, we assume that there are two neighboring segments each of duration of 2 s, shifted every 20 ms. The feature values for the 5 features are computed for 25 segment shifts (which is equal to 1 s resolution). Features are assumed to follow the Gaussian distribution. The covariance matrices, $\mathbf{X}$ for the first segment and $\mathbf{Y}$ for the second segment, are calculated. The proposed statistical measure, K, is a combination of the arithmetic mean $a(\mathbf{X},\mathbf{Y})$, the geometric mean $g(\mathbf{X},\mathbf{Y})$, and the harmonic mean $h(\mathbf{X},\mathbf{Y})$ of the eigenvalues of $\mathbf{Y}\mathbf{X}^{-1}$. The aforementioned means definitions can be found in [8]. We employ $\log(a(\mathbf{X},\mathbf{Y})^2/g(\mathbf{X},\mathbf{Y})h(\mathbf{X},\mathbf{Y}))$. Symmetrization is expected to improve the classification performance, compared to both asymmetric terms taken individually in [1], [8]. Symmetrization results to:

$$\begin{aligned} K &= \log(a(\mathbf{X},\mathbf{Y})^2/g(\mathbf{X},\mathbf{Y})h(\mathbf{X},\mathbf{Y})) + \\ &\quad \log(a(\mathbf{Y},\mathbf{X})^2/g(\mathbf{Y},\mathbf{X})h(\mathbf{Y},\mathbf{X})) \\ &= 3\log\mathrm{tr}(\mathbf{X}\mathbf{Y}^{-1}) + 3\log\mathrm{tr}(\mathbf{Y}\mathbf{X}^{-1}) - 6\log\delta \end{aligned} \quad (2)$$

where δ=50. Next, we compare K with an ad hoc threshold $\widetilde{\vartheta}$. If $K > \widetilde{\vartheta}$ then a turn point is detected between the two segments, otherwise the potential speaker change point is discarded.

In the second module, the first segment is modeled by the Gaussian distribution $\mathcal{N}(\mathbf{m}_X, \mathbf{\Sigma}_X)$, the second segment by $\mathcal{N}(\mathbf{m}_Y, \mathbf{\Sigma}_Y)$ and the union of the two segments by $\mathcal{N}(\mathbf{m}_Z, \mathbf{\Sigma}_Z)$. The extracted features are the MFCCs and delta MFCCs. The Euclidean distance and the T^2 Hotelling statistic. T^2 Hotelling statistic between $\mathbf{m}_X$ and $\mathbf{m}_Y$ were used. T^2 Hotelling statistic is defined as [5]:

$$d_{T^2}(\mathbf{m}_X, \mathbf{m}_Y) = \frac{N_X N_Y}{N_X + N_Y}(\mathbf{m}_X - \mathbf{m}_Y)^T \mathbf{\Sigma}_Z^{-1}(\mathbf{m}_X - \mathbf{m}_Y) \quad (3)$$

where N_X, N_Y is the number of frames within each segment respectively and each frame has a duration of 40 ms. In this case, a tandem Bayesian Network is utilized, since in the two detector case, the tandem network is dominant [9]. As a result, we examine every potential speaker change point by using the Euclidean distance between $\mathbf{m}_X$ and $\mathbf{m}_Y$ and then, we re-examine them using T^2 Hotelling statistic $d_{T^2}(\mathbf{m}_X, \mathbf{m}_Y)$. The reason for using Euclidean distance first in the chain of the Bayesian Network is that it has been proven to yield poorer results that T^2 Hotelling statistic experimentally, which is explained by the fact that Euclidean distance does not take into account the correlation of the data, since it disregards $\mathbf{\Sigma}_Z$. In [9], it is established that it is better to put the best detector later in the Bayesian Network chain.

BIC is applied in the third module. BIC is applied last, because it performs better when the segments are long enough [1], [11]. BIC is computed in conjunction with the MFCCs and the potential sets of potential speaker change points are fed to BIC, used with delta MFCCs to yield the final set of speaker change points.

C. The third system

The input audio stream is first down-sampled to 10kHz, 16 bits, mono channel format. The speech stream is then pre-emphasized using the filter $H(z) = 1 - 0.97z^{-1}$ and is divided into 25 ms analysis frames with 5 ms overlap. Non speech and silence frames are removed [10]. The system considers only voiced frames, from which 10-order LSP features are extracted, that are assumed to be Gaussian distributed. Afterwards, speech segments are formed by accumulating the necessary number of voiced frames, so that there are sufficient data to prevent the ill-posed LSP covariance matrices. This implies that each segment should at least include 55 voiced frames, which corresponds to a minimum segment duration of

1.375 s. The sliding of the segment window is 0.5 s of voiced speech. Thus, the basic processing unit is a 1.375 s segment and the temporal resolution of the segmentation is 0.5 s.

First, speaker change detection is coarsely performed using a metric-based approach to calculate the distance between consecutive speech segments. Each speech segment is represented by a multivariate Gaussian $\mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$. The Kullback-Leibler (K-L) divergence is used to estimate the distance between two speech segments i and j, defined as $D(i,j) = \frac{1}{2}tr[(\mathbf{\Sigma}_i - \mathbf{\Sigma}_j)\ (\mathbf{\Sigma}_i^{-1} - \mathbf{\Sigma}_j^{-1})]$ [11]. A potential speaker turn point is detected between two consecutive speech segments whenever the following three conditions are verified: $D(i,i+1) > D(i+1,i+2),\ D(i,i+1) > D(i-1,i)$, and $D(i,i+1) > \theta_i$ [11]. The first two conditions guarantee that a local maximum exists. The third condition assures that the prominence of the distance peak is high enough to be considered relevant. However, it is based on a threshold, whose value is not set trivially: a too high value would imply a high miss detection rate, and a too low value would increase the false alarm rate. Lu and Zhang proposed an automatic data-dependent threshold setting method [11]:

$$\theta_i = \alpha\ \frac{1}{N} \sum_{n=0}^{N} D(i-n-1, i-n) \tag{4}$$

where N is the number of the past speech segments used for threshold estimation. We set $N = 3$ and $\alpha = 0.4$. In order to reduce the false alarm rate, we use the BIC difference between speech segments $\mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1)$ and $\mathcal{N}(\mathbf{m}_2, \mathbf{\Sigma}_2)$, defined as:

$$\begin{aligned} BIC(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) = \frac{1}{2}\ ((N_1+N_2)\log|\mathbf{\Sigma}| - N_1 \log|\mathbf{\Sigma}_1| \\ -N_2 \log \mathbf{\Sigma}_2)\ - \frac{1}{2}\ \lambda(\delta + \frac{1}{2}\delta(\delta+1))\ \log(N_1+N_2) \end{aligned} \tag{5}$$

where N_1 and N_2 are the number of feature vectors used to estimate $\mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1)$ and $\mathcal{N}(\mathbf{m}_2, \mathbf{\Sigma}_2)$, respectively, $\mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$ is a single Gaussian estimated using both speech segments, and $\lambda = 1$. If $BIC(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$ admits a positive value, the two speech segments are likely to originate from different speakers, so the speaker change point is accepted. Otherwise, no speaker change point is declared.

Next, we utilize the approach proposed in [11], in order to refine the detected speaker change points and build the speaker models. Speaker models are stored using a quasi-GMM approach, but we propose a different implementation of the quasi-GMM procedure. Each speaker is modeled by a quasi-GMM with at most 32 Gaussian components. In our implementation, if no speaker change is detected at a specific point, instead of discarding the arriving speaker data when the model reaches a number of 32 time-dependent Gaussian components, the oldest component is marked in the current speaker model and is replaced by the component created from the new speech data. This mechanism is robust to speakers whose voice starts to present changes after talking for long periods of time, or to long-term changes of background noise or recording conditions.

Let us denote by $\mathcal{N}(\mathbf{m}_2, \mathbf{\Sigma}_2)$ the current speech segment and by $\mathcal{N}(\mathbf{m}_{1j}, \mathbf{\Sigma}_{1j})$ the current quasi-GMM speaker model having S Gaussian densities with N_{1j} feature vectors each. The distance between $\mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1)$ and $\mathcal{N}(\mathbf{m}_2, \mathbf{\Sigma}_2)$ is estimated as:

$$D = \sum_{j=1}^{S} w_{ij}\ BIC(\mathbf{\Sigma}_{1j}, \mathbf{\Sigma}_2) \tag{6}$$

where $w_{ij} = \frac{N_{1j}}{N_1}$ and $N_1 = \sum_{j=1}^{S} N_{1j}$. If $D > 0$, the potential speaker change point detected in the first stage is confirmed as a real speaker boundary by the BIC refinement procedure, otherwise it is rejected.

III. Experimental results

In order to assess the performance of the proposed algorithm, the TIMIT database was used [12]. A total of 15 recordings is created by concatenating speakers from the TIMIT database. 10 randomly selected recordings have been used to evaluate the performance of the three aforementioned systems, while the remaining 5 recordings have been used for tuning the respective system parameters. Examples of such parameters are ϑ in the first system, K in the second system, and α in the third system.

Two sets of figures of merit are commonly used. On the one hand, one may use the false alarm rate (FAR) and the miss detection rate (MDR) defined as:

$$FAR = \frac{FA}{GT + FA} \quad MDR = \frac{MD}{GT} \tag{7}$$

where FA denotes the number of false alarms, MD the number of miss detections, and GT stands for the actual number of speaker turns, i.e. the ground truth. A false alarm occurs when a speaker turn is detected although it does not exist, a miss detection MD occurs when the process does not detect an existing speaker turn. On the other hand, one may employ the precision (PRC), recall (RCL),and F_1 rates given by:

$$PRC = \frac{CFC}{DET} \quad RCL = \frac{CFC}{GT} \quad F_1 = \frac{2\ PRC\ RCL}{PRC + RCL} \tag{8}$$

where CFC denotes the number of correctly found changes and DET is the number of the detected speaker changes. F_1 admits a value between 0 and 1. The higher its value is, the better performance is obtained. Between the pairs (FAR, MDR) and (PRC, RCL) the following relationships hold $MDR = 1 - RCL$ and $FAR = \frac{RCL\ FA}{DET\ PRC + RCL\ FA}$. Table I demonstrates the performance of the 3 systems in terms of mean value and standard deviation for all the five features of merit over the 10 randomly selected recordings, created by concatenating speakers from the TIMIT database.

IV. Statistical Analysis

Our aim is to test whether the performance of each system differentiates significantly from the other, with respect to the mean of all the five figures of merit. One-way ANOVA is selected for this purpose and is applied for a 95% confidence level. The null hypothesis, that the groups means are equal is tested, i.e. the systems do not differentiate significantly from

TABLE I

PERFORMANCE OF THE 3 SYSTEMS ON 10 RANDOMLY SELECTED RECORDINGS IN TERMS OF MEAN VALUE AND STANDARD DEVIATION.

System examined	First System	Second System	Third System
PRC **(**mean**)**	0.780	0.490	0.678
PRC **(**st. dev.**)**	0.137	0.040	0.155
RCL **(**mean**)**	0.700	0.812	0.546
RCL **(**st. dev.**)**	0.136	0.111	0.108
F_1 **(**mean**)**	0.720	0.607	0.580
F_1 **(**st. dev.**)**	0.081	0.037	0.11
FAR **(**mean**)**	0.218	0.455	0.244
FAR **(**st. dev.**)**	0.135	0.060	0.117
MDR **(**mean**)**	0.305	0.188	0.444
MDR **(**st. dev.**)**	0.136	0.111	0.108

one another. The alternative hypothesis states that the groups means are unequal, i.e. at least one of the systems differs from the rest. The F-statistic value and the p-value for all five efficiency measures is shown in Table II. As it is easy to deduct from Table II, the three systems are statistically different, with respect to PRC, RCL, F_1, and MDR, but there appears to be no statistically significant difference for FAR.

TABLE II

F-STATISTIC VALUES AND p-VALUES FOR PRC, RCL, F_1, FAR, AND MDR OF THE 3 SYSTEMS TESTED ON THE SAME RECORDINGS.

	PRC	RCL	F_1	FAR	MDR
F**-statistic**	14.5367	9.8309	8.2102	15.3861	11.5049
p**-value**	0.00005	0.00006	0.0016	0.00003	0.00024

However, no information is provided by ANOVA about which system differentiates. There are post hoc, or posterior tests to tackle this problem. Tukey's method, or honestly significant differences method, is employed. Tukey's method makes all pair-wise comparisons of means while maintaining the confidence level at a pre-defined value, which in this case is set to 95%. Tukey's method is applied for all figures of merit. The corresponding confidence intervals of all pair-wise comparisons of the three systems for the five figures of merit can be seen in Table III. If the confidence interval includes zero, the differences are not significant. Accordingly, the differences are significant, if zero does not belong to the confidence interval. As it can be easily deduced from Table III, the first and second system cannot be considered equivalent with respect to PRC, F_1 and FAR. But they are of equal performance with respect to RCL and MDR. Similarly, the first and third system present distinct F_1 and MDR, while for PRC, RCL, and FAR their efficiency can be considered to be the same. Finally, the comparison between second and third systems, reveals that F_1 values are not significantly different, whereas PRC, RCL, FAR, and MDR values are.

V. CONCLUSIONS

Three automatic speaker segmentation BIC-based systems are described. Each system was evaluated on 10 randomly selected recordings and the results have been statistically analyzed by one-way ANOVA and post hoc Tukey test. The first system appears to favors accuracy, as can be deducted from its lowest FAR score. The second system exhibits the lower MDR, which is consistent with the research community's trend to consider MDR more important than FAR, since the latter is easier to remedy. The third system puts a higher emphasis on the real-time operation and also manages to maintain low FAR levels.

TABLE III

95% CONFIDENCE INTERVALS FOR ALL PAIR-WISE COMPARISONS OF THE 3 SYSTEMS FOR PRC, RCL, F_1, FAR, AND MDR.

Systems	PRC	RCL	F_1	FAR	MDR
1^{st} - 2^{nd}	[0.16,0.43]	[-0.26,0.02]	[0.02,0.20]	[-0.36,-0.12]	[-0.02,0.25]
1^{st} - 3^{rd}	[-0.03,0.24]	[-0.01,0.27]	[0.05,0.23]	[-0.13,0.12]	[-0.28,-0.01]
2^{nd} - 3^{rd}	[-0.32,-0.05]	[0.11,0.39]	[-0.06,0.12]	[0.11,0.35]	[-0.39,-0.12]

ACKNOWLEDGMENT

V. Moschou acknowledges the FP6 European Union Network of Excellence MUSCLE "Multimedia Understanding through Semantics, Computation and Learning" (FP6-507752). M. Kotti acknowledges the support project 03ED 849 co-funded by the European Union and the Greek Secretariat of Research and Technology (Hellenic Ministry of Development) of the Operational Program for Competitiveness within the 3rd Community Support framework.

REFERENCES

[1] P. Delacourt and C. J. Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing," *Speech Communication*, vol. 32, pp. 111-126, September 2000.

[2] J. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," *IEEE Signal Processing Letters*, vol. 11, no. 8, pp. 649-651, August 2004.

[3] H. Kim, D. Elter, and T. Sikora, "Hybrid speaker-based segmentation system using model-level clustering," in *Proc. 2005 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, pp. 745-748, Philadelphia, USA, March 2005.

[4] M. Kotti, E. Benetos, and C. Kotropoulos, "Automatic speaker change detection with the Bayesian Information Criterion using MPEG-7 features and a fusion scheme", in *Proc. IEEE Int. Symposium Circuits and Systems*, 2006.

[5] R. Huang and J. H. L. Hansen, "Advances in unsupervised audio segmentation for the broadcast news and ngsw corpora," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, pp. 741-744, May, 2004.

[6] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems", *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 5. pp.1557-1565, 2006.

[7] F. Van der Heijden, R. P. W. Duin, D. de Ridder, and D. M. J. Tax, *Classification, parameter estimation and state estimation: An engineering approach using MATLAB*, London, UK: Wiley, 2004.

[8] F. Bimbot, I. Magrin-Chagnolleau, and L. Mathan, "Second-order statistical measures for text-independent speaker identification," *Speech Communication*, vol. 17, pp. 177-192, 1995.

[9] Y. Zhu and X. Rong, "Unified fusion rules for multisensor multihypothesis network decision systems", *IEEE Trans. Systems, Man, and Cybernetics*, vol. 33, no.4, pp. 502-513, 2003.

[10] D. G. Childers, *Speech processing and synthesis toolboxes*, London, UK: Wiley, 1999.

[11] L. Lu and L. Zhang, "Real-time unsupervised speaker change detection.", in *Proc. 16th Int. Conf. Pattern Recognition*, vol. 2 pp. 358-361, 2002.

[12] J. S. Garofolo, "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus," Linguistic Data Consortium, Philadelphia, 1993.