AUTOMATIC SPEAKER SEGMENTATION USING MULTIPLE FEATURES AND DISTANCE MEASURES: A COMPARISON OF THREE APPROACHES

Margarita Kotti[1], *Luís Gustavo P. M. Martins*[2], *Emmanouil Benetos*[1], *Jaime S. Cardoso*[2], *Constantine Kotropoulos*[1*]

[1]Department of Informatics, Aristotle Univ. of Thessaloniki
Box 451, Thessaloniki 541 24, Greece
E-mail: {mkotti,empeneto,costas}@zeus.csd.auth.gr

[2]INESC Porto, Porto
Portugal
E-mail: {lmartins,jsc}@inescporto.pt

ABSTRACT

This paper addresses the problem of unsupervised speaker change detection. Three systems based on the Bayesian Information Criterion (BIC) are tested. The first system investigates the AudioSpectrumCentroid and the AudioWaveformEnvelope features, implements a dynamic thresholding followed by a fusion scheme, and finally applies BIC. The second method is a real-time one that uses a metric-based approach employing the line spectral pairs and the BIC to validate a potential speaker change point. The third method consists of three modules. In the first module, a measure based on second-order statistics is used; in the second module, the Euclidean distance and T^2 Hotelling statistic are applied; and in the third module, the BIC is utilized. The experiments are carried out on a dataset created by concatenating speakers from the TIMIT database, that is referred to as the TIMIT data set. A comparison between the performance of the three systems is made based on t-statistics.

1. INTRODUCTION

Automatic speech segmentation aims at finding the speaker change points in an audio stream. It is a preprocessing task for audio indexing, speaker identification - verification - tracking, automatic transcription, information extraction, topic detection, speech summarization and retrieval.

Among the most popular techniques for speaker segmentation are these based on the Bayesian Information Criterion (BIC) [1, 2, 3, 6, 7, 9]. However, BIC-based speaker segmentation is time consuming, a fact that has motivated research towards alleviating its computational demands [8]. Tritschler and Gopinnath used the BIC on the mel-cepstrum coefficients (MFCCs) [3]. Delacourt and Wellekens proposed a two-pass segmentation technique called DISTBIC that improved the performance by utilizing distance-based segmentation before applying the BIC. [1]. Ajmera et al. introduced a BIC alternative, which does not need tuning [2]. Meanwhile, novel features like the smoothed zero crossing rate (SZCR), the perceptual minimum variance distortionless response, and the filterbank log coefficients were introduced by Huang and Hansen [8]. METRIC-SEQDAC is another method [7]. Finally, a hybrid algorithm, which combines metric-based segmentation with the BIC and model-based segmentation with Hidden Markov Models (HMMs) is proposed [6].

The major contribution of this paper is in the comparative performance of three speaker segmentation systems. All systems are based on the BIC and their efficiency is tested on the TIMIT dataset using the same experimental protocol. Moreover, their performance is further assessed by using t-statistics. The first system investigates scalar and vector features as described in Section 2, an adaptive dynamic thresholding for the scalar features, and a fusion scheme which combines the partial results so as to achieve a better performance than that obtained without fusion. The second system is an improved real-time speaker change detection system able to recognize speaker turn points with the shortest possible delay, without having access to the entire speech stream. This scenario imposes certain limitations on the computational load of the algorithm [4, 5]. In this system, the processing has two main stages: In the first stage, a metric-based approach is implemented using the Line Spectral Pairs (LSP). In the second stage, the BIC is used to validate the potential speaker change points detected previously. In the third novel system, there are three modules: In the first module, scalar features and a second-order statistical measure is implemented. In the second module, the MFCCs and the delta MFCCs are used in conjunction with the T^2 Hotelling statistic and the Euclidean distance. Finally, in the last module the MFCCs and the delta MFCCs are utilized, but the decisions are taken with respect to the BIC.

The rest of the paper is organized as follows. In Section 2, the three systems are described. Experimental results are shown in Section 3 and conclusions are drawn in Section 4.

2. THREE SYSTEMS FOR SPEAKER CHANGE DETECTION

2.1. The first system

The system relies on the BIC variant proposed in [2]. The selection of the appropriate features is of great importance, since the accurate representation of the audio signal is vital. We utilize the MFCCs, the maximum magnitude of the DFT coefficients in a speech frame, the short-time energy (STE), the AudioSpectrumCentroid, and the AudioWaveformEnvelope. Multiple passes are allowed. In the first four passes, we use the MFCCs; in the fifth pass the maximum DFT magnitude; in the sixth pass the STE; in the seventh pass the MFCCs; in the eighth pass the AudioSpectrumCentroid; in the ninth pass the maximum DFT magnitude, and in the last pass the AudioWaveformEnvelope. Multiple passes are employed because after each pass, the number of chunks is decreased, due to specific potential change points are discarded being false. Several researchers have come to the conclusion that the larger the chunks are, the better the performance is, because there is enough data for satisfactory parameter estimation of the speaker model [1, 3, 4, 5, 8]. The decisions taken in one pass are fed to the next pass as in a Bayesian network.

Every speaker is represented with a multivariate Gaussian probability density function (pdf) with mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. The pdf parameters are automatically updated when more data are available. Utilizing the fact that the chunks are becoming larger, we employ a constant updating of the speaker models [4, 5, 8].

*This work has been supported by the FP6 European Union Network of Excellence MUSCLE "Multimedia Understanding through Semantics, Computation and Learning" (FP6-507752).

The dynamic thresholding refers only to scalar features such as: the maximum magnitude of DFT, the STE and the AudioWaveformEnvelope. We start with an ad hoc threshold ϑ that is determined after a considerable number of experiments during which we compute the F_1 measure, as defined in (15), for several threshold values and then we retain the value which maximizes the F_1 measure. Let us consider a recording that has I chunks and $I-1$ possible speaker change points. The value of I is determined at the previous pass. 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 $f(k)$ and $f(k+1)$ 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:

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

Whenever a feature vector is employed (such as the MFCCs and the AudioSpectrumCentroid) the BIC is used. To estimate the GMM needed in the BIC, the EM algorithm is used. The EM algorithm may converge at local minima. There is no guarantee that a local minimum coincides with the global minimum or that there is only one local minimum. This issue, combined with the fact that the BIC is a weak classifier leads us to propose a fusion scheme. Thus, we could theoretically reduce the error introduced by the EM algorithm by repeating the experiment multiple times, say R times, and applying a majority voting. To be more specific, for each repetition we obtain a set of possible speaker turn points. The set of change points after all repetition in this pass consists of those potential speaker change points that make their appearance at a sufficient frequency S. Both R and S are determined heuristically. Typical values for R and S are 5 and 4, respectively. The just described procedure is detailed in [9].

2.2. The second system

This system is an improved version of the real-time speaker change detection system described in [9]. The second system starts by downsampling the input speech audio to 8 kHz, 16 bits mono channel format and applying pre-emphasis. The speech stream is then divided into analysis frames of 25ms duration without overlap. From each frame 10-order LSPs features are extracted [9]. In contrast to the system described in [9], where only the voiced speech frames determined by a voiced/unvoiced/silence classifier were processed, the current system processes all the available frames. Although one would expect the system performance to deteriorate by omitted the voiced/unvoiced/silence classifier, the experimental results demonstrate the opposite. Such findings may be attributed to the limited success of the aforementioned classifier to determine accurately the voiced frames.

In the first stage, speaker change detection is coarsely performed using a metric-based approach to calculate the distance between consecutive and non-overlapping speech segments. Each speech segment includes 55 speech frames corresponding to 1.375 sec. Assuming that the LSPs are Gaussian distributed, each speech segment can be modelled by a multivariate Gaussian. It has been found by experiments that the aforementioned number of frames is the minimum number of frames that prevents an ill-conditioned covariance matrix for 10th -order LSPs. The Kullback-Leibler (K-L) divergence shape distance [12] is used to estimate the distance between two subsequent speech segments i and j:

$$D(i,j) = \frac{1}{2}\mathrm{tr}[(\mathbf{\Sigma}_i - \mathbf{\Sigma}_j)(\mathbf{\Sigma}_j^{-1} - \mathbf{\Sigma}_i^{-1})] \quad (2)$$

where tr stands for the trace operator.

Let $\mathbf{f}_n$ denote the 10th-order LSP feature vector extracted from the nth frame. The covariance matrix $\mathbf{\Sigma}_i$ of the ith segment is:

$$\mathbf{\Sigma}_i = \begin{cases} \mathrm{cov}\{\mathbf{f}_{1+27(\frac{i-1}{2})}, \ldots, \mathbf{f}_{55+27(\frac{i-1}{2})}\} & \text{when } i \text{ is odd} \\ \mathrm{cov}\{\mathbf{f}_{56+27(\frac{i}{2}-1)}, \ldots, \mathbf{f}_{110+27(\frac{i}{2}-1)}\} & \text{when } i \text{ is even} \end{cases} \quad (3)$$

where cov stands for the covariance matrix operator. Using (2), a potential speaker turn point can be detected between two segments, whenever the following conditions are satisfied:

$$D(i,i+1) > D(i+2,i+3) \quad (4)$$

$$D(i,i+1) > D(i-2,i-1) \quad (5)$$

$$D(i,i+1) > \vartheta_i^{''}. \quad (6)$$

The first two conditions guarantee that a local maximum exists. The third condition assures that the local maximum is a prominent one. However, it is based on a threshold $\vartheta_i^{''}$, whose value cannot be selected trivially. An automatic threshold selection, which is based on the values of the distances between the past N consecutive speech segments, is proposed in [4]:

$$\vartheta_i^{''} = \frac{\alpha}{N}\sum_{n=1}^{N} D(i-2n, i+1-2n) \quad (7)$$

where α is a scaling factor used to tune the system response. The best system performance is obtained for α=0.8 when N=3.

To reduce the false alarm rate, the BIC is used in the second stage to validate any potential speaker change point detected by the coarse segmentation procedure. Let $\mathcal{N}(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i)$ and $\mathcal{N}(\boldsymbol{\mu}_j, \mathbf{\Sigma}_j)$ be the Gaussian models derived from two speech segments and N_i and N_j be the corresponding number of feature vectors used in the estimation. Let also $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ be a single Gaussian model estimated on the union of the aforementioned speech segments, the BIC difference between the two models can be defined as:

$$BIC(\mathbf{\Sigma}_i, \mathbf{\Sigma}_j) = \frac{1}{2}\big((N_i+N_j)\log|\mathbf{\Sigma}| - N_i\log|\mathbf{\Sigma}_i| - N_j\log|\mathbf{\Sigma}_j|\big) - \frac{1}{2}\lambda(\delta + \frac{1}{2}\delta(\delta+1))\log(N_i+N_j) \quad (8)$$

where λ is the penalty factor for the model complexity, here set equal to 0.6, and δ is the feature vector dimension (i.e. $\delta = 10$). If $BIC(\mathbf{\Sigma}_i, \mathbf{\Sigma}_j)$ takes a positive value, the two speech segments are likely to originate from different speakers, so the speaker change point is accepted. Otherwise, the two segments correspond to the same speaker and no speaker change point is declared.

To implement a continuous updating of the speaker model, we make use of a solution based on quasi-GMM modelling, a non-iterative technique that allows real-time operation with a reasonable accuracy [4]. Instead of considering one audio segment modelled by $\mathcal{N}(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i)$, a speaker model is used which is approximated by a quasi-GMM $\mathcal{N}(\boldsymbol{\mu}_{qGMM}, \mathbf{\Sigma}_{qGMM})$. This speaker model is composed of S Gaussian mixtures $\mathcal{N}(\boldsymbol{\mu}_m, \mathbf{\Sigma}_m)$, $m = 1, 2, \ldots, S$ over N_m feature vectors each, for S=32 [4]. As new speech data arrive, only the data from the odd-numbered speech segments are used for

updating the current quasi-GMM speaker model. This allows better statistical independence between the models under comparison as required by the BIC. Let also $\mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, with $j = i + 1$ be the Gaussian density of the speech segment located just after the potential speaker change point under validation. The distance between the density models $\mathcal{N}(\boldsymbol{\mu}_{qGMM}, \boldsymbol{\Sigma}_{qGMM})$ and $\mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ can be roughly estimated as:

$$BIC(\boldsymbol{\Sigma}_{qGMM}, \boldsymbol{\Sigma}_j) = \sum_{m=1}^{S} w_m \, BIC(\boldsymbol{\Sigma}_m, \boldsymbol{\Sigma}_j) \tag{9}$$

where $w_m = \frac{N_m}{N_{qGMM}}$ and $N_{qGMM} = \sum_{m=1}^{S} N_m$. Whenever $BIC(\boldsymbol{\Sigma}_{qGMM}, \boldsymbol{\Sigma}_j) > 0$, the potential speaker change previously detected by the metric-based approach is confirmed as an actual speaker boundary by the BIC refinement procedure.

2.3. The third system

The third system is a novel one that consumes less time than the first system, because it avoids the multiple iterations. It resembles the second system. To be more specific, it consists of three modules. In the first module, the best 5 scalar features among 24 features are derived and a second-order statistical measure is employed. In the second module, the MFCCs are utilized in conjunction with the Euclidean distance and the T^2 Hotelling statistic. In the third module, MFCCs and delta MFCCs are used in combination with the BIC.

In the first module, we investigate a total set of 24 features. The set includes the mean and the variance of the following feature values, their first-order (delta) and second-order (delta-delta) differences: magnitude of the DFT, STE, AudioWaveformEnvelope, and the maximum of AudioSpectrumCentroid. A feature selection algorithm is applied in order to derive the optimum feature subset for classification. In particular, we select the best 5 out of the 24 features. The search strategy implemented is Branch and Bound. The criterion utilized for selection is based on the maximization of the ratio of the trace of the inter class scatter matrix over the trace of the intra class scatter matrix [11]. Starting from the most effective feature, the 5 best selected features are: the mean magnitude of the DFT, the delta AudioWaveformEnvelope, the mean STE, the AudioWaveformEnvelope, and finally the variance of the delta magnitude of the DFT.

Next, we consider each feature among the 5 best sequentially. Each audio recording is segmented to windows with duration of 2 sec and the feature values are computed for 2 adjacent windows. Each window is divided into acoustic vectors of 0.040 sec duration with overlap of 0.020 sec. For every acoustic vector we compute a scalar feature. For the first window, the sequence of feature values has a covariance matrix denoted by $\mathbf{X}$. Let $\mathbf{Y}$ be the covariance matrix of the sequence of feature values for the second window. The statistical measure proposed is a second-order statistical measure value is computed for this window pair. It 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 $\mathbf{Y}\mathbf{X}^{-1}$ defined as follows:

$$a(\mathbf{X}, \mathbf{Y}) = \frac{\mathrm{tr}(\mathbf{Y}\mathbf{X}^{-1})}{\delta} \quad , \quad g(\mathbf{X}, \mathbf{Y}) = \left(\frac{\det(\mathbf{Y})}{\det(\mathbf{X})} \right)^{\frac{1}{\delta}}$$
$$h(\mathbf{X}, \mathbf{Y}) = \frac{\delta}{\mathrm{tr}(\mathbf{X}\mathbf{Y}^{-1})} \tag{10}$$

The ratio $\log \frac{a(\mathbf{X},\mathbf{Y})}{g(\mathbf{X},\mathbf{Y})}$ has been previously used in [1, 13] and the ratio $\log \frac{a(\mathbf{X},\mathbf{Y})}{h(\mathbf{X},\mathbf{Y})}$ was proposed in [13]. With the combination of these two we employe $\log(a(\mathbf{X}, \mathbf{Y})^2 / g(\mathbf{X}, \mathbf{Y}) h(\mathbf{X}, \mathbf{Y}))$ and we expect improved results. Moreover, symmetrization can improve the classification performance, compared to both asymmetric terms taken individually [1, 13]. Symmetrization results to the proposed statistic measure:

$$\begin{aligned} K &= \log(a(\mathbf{X}, \mathbf{Y})^2 / g(\mathbf{X}, \mathbf{Y}) h(\mathbf{X}, \mathbf{Y})) + \\ &\quad \log(a(\mathbf{Y}, \mathbf{X})^2 / g(\mathbf{Y}, \mathbf{X}) h(\mathbf{Y}, \mathbf{X})) \\ &= 3 \log \mathrm{tr}(\mathbf{X}\mathbf{Y}^{-1}) + 3 \log \mathrm{tr}(\mathbf{Y}\mathbf{X}^{-1}) - 6 \log \delta \end{aligned} \tag{11}$$

where δ=50. Next, we compare K with an ad hoc threshold $\widetilde{\vartheta}$. If $K > \widetilde{\vartheta}$ then a turn point is assumed between the two windows, else the potential speaker change point is discarded. Then the pair of the windows is shifted by 0.5 sec. This procedure is repeated for every scalar feature so 5 different possible speaker change points sets are computed, each for every feature. These sets are fused in a parallel Bayesian Network so as to produce the final set of possible speaker change points. A parallel network was used because it outperforms a tandem network while detecting Gaussian signal in Gaussian noise [10].

In the second module, the implemented features are the MFCCs and the utilized distances are the Euclidean distance and the T^2 Hotelling statistic. If the first window is modelled by the Gaussian distribution $\mathcal{N}(\mathbf{m}_X, \boldsymbol{\Sigma}_X)$, the second window by $\mathcal{N}(\mathbf{m}_Y, \boldsymbol{\Sigma}_Y)$ and the union of the two windows by $\mathcal{N}(\mathbf{m}_Z, \boldsymbol{\Sigma}_Z)$, T^2 Hotelling statistic is defined as [8]:

$$d_{T_2} = \frac{N_X N_Y}{N_X + N_Y} (\mathbf{m}_X - \mathbf{m}_Y)^T \boldsymbol{\Sigma}_Z^{-1} (\mathbf{m}_X - \mathbf{m}_Y) \tag{12}$$

where N_X, N_Y is the number of frames within each window respectively and each frame has a duration of 40 msec. In this case, a tandem Bayesian Network is utilized, since in the two detector case the tandem network is dominant [10]. It has also been proven that it is better to place the detector with the better performance later in the chain. We experimentally found that the T^2 Hotelling statistic outperforms the Euclidean distance, which is rather logical, since the Euclidean distance does not take into account the correlation that might exist between the data of the first and second window by not taking into account the $\boldsymbol{\Sigma}_Z$ matrix. As a result we first examine the potential speaker change points by using the Euclidean distance and then, we re-examine them using T^2 Hotelling statistic.

In the final third module, the BIC is implemented. The reason why BIC is the applied last is the fact that BIC performs better when the segments are long enough [1, 3, 4, 5]. This module has two stages. In the first stage, the BIC is computed in conjunction with the MFCCs and the potential sets of potential speaker change points are fed to the second stage , where BIC is used with delta MFCCs to produce the final set of speaker change points.

3. EXPERIMENTAL RESULTS

In order to assess the performance of the aforementioned algorithms the TIMIT dataset was created by concatenating speakers from the TIMIT database. TIMIT is an acoustic-phonetic database including 6300 sentences and 630 speakers who speak English. The audio format is PCM, the audio samples are quantized in 16 bit, the recordings are single-channel, the mean duration is 3.28 sec and the standard deviation (st. dev.) is 1.52 sec. For all three systems parameters were fine-tuned using the complete TIMIT dataset (43 speech files), not including any of the files used on the evaluation.

Two pairs of figures of merit are used to assess the performance of a speaker change detection system. On the one hand, one may

use the false alarm rate (FAR) and the miss detection rate (MDR) defined as:

$$FAR = \frac{FA}{GT + FA} \quad MDR = \frac{MD}{GT} \tag{13}$$

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) and recall (RCL) rates given by:

$$PRC = \frac{CFC}{DET} \quad RCL = \frac{CFC}{GT} \tag{14}$$

where CFC denotes the number of correctly found changes and DET is the number of the detected speaker changes. For the latter pair, another objective figure of merit is the F_1 measure

$$F_1 = \frac{2\,PRC\,RCL}{PRC + RCL} \tag{15}$$

that 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:

$$MDR = 1 - RCL \quad FAR = \frac{RCLFA}{DETPRC + RCLFA}. \tag{16}$$

Table 1 demonstrates the performance for 10 randomly selected test recordings extracted from TIMIT database not included in the training procedure. In Table 2, the results for the same 10 randomly selected test recordings are demonstrated for the second system, whilst in Table 3 the corresponding results are shown for the third system. The efficiency has been presumed dropping whenever the speaker's utterance has a duration of less than 1-2 sec, as it was expected [1, 3, 4].

Table 1. Performance of the first system on the TIMIT dataset.

Index	PRC	RCL	F_1	FAR	MDR
1	0.83	0.56	0.67	0.17	0.44
2	0.90	0.75	0.82	0.10	0.25
3	1.00	0.45	0.62	0.0	0.55
4	0.62	0.82	0.72	0.36	0.18
5	0.64	0.90	0.75	0.36	0.10
6	0.85	0.79	0.81	0.15	0.21
7	0.69	0.65	0.67	0.31	0.35
8	0.93	0.76	0.84	0.07	0.24
9	0.69	0.58	0.63	0.31	0.42
10	0.65	0.69	0.67	0.35	0.31
mean	0.780	0.700	0.720	0.218	0.305
st. dev.	0.137	0.136	0.008	0.135	0.136

Table 2. Performance of the second system on the TIMIT dataset.

Index	PRC	RCL	F_1	FAR	MDR
1	0.67	0.89	0.76	0.31	0.11
2	0.60	0.90	0.72	0.38	0.10
3	0.78	0.64	0.70	0.15	0.36
4	0.78	0.78	0.78	0.18	0.22
5	0.41	0.78	0.54	0.53	0.22
6	0.61	0.85	0.71	0.35	0.15
7	0.73	0.85	0.79	0.24	0.15
8	0.72	0.76	0.74	0.23	0.24
9	0.74	0.82	0.78	0.23	0.18
10	0.78	0.74	0.76	0.17	0.26
mean	0.680	0.80	0.730	0.280	0.200
st. dev.	0.120	0.080	0.070	0.120	0.080

Table 3. Performance of the third system on the TIMIT dataset.

Index	PRC	RCL	F_1	FAR	MDR
1	0.45	1.00	0.62	0.55	0.00
2	0.50	0.58	0.54	0.37	0.42
3	0.53	0.82	0.64	0.42	0.18
4	0.47	0.82	0.60	0.48	0.18
5	0.47	0.80	0.59	0.47	0.20
6	0.50	0.79	0.61	0.44	0.21
7	0.58	0.82	0.68	0.37	0.18
8	0.46	0.94	0.62	0.53	0.06
9	0.48	0.74	0.58	0.44	0.26
10	0.46	0.81	0.59	0.48	0.19
mean	0.490	0.812	0.607	0.455	0.188
st. dev.	0.040	0.111	0.037	0.060	0.111

4. CONCLUSIONS

The performance of three BIC-based speaker segmentation systems is compared in this paper. Each system was evaluated on the TIMIT dataset. The first system puts a higher emphasis on the accuracy than the real-time operation and is the most stable of all since the standard deviation of F_1 is 0.008. The second system favors the real-time operation. The third systems tries to compensate between the first system and the second one. The third system is also more stable than the second one since the standard deviation of F_1 is approximately 50% of that of the second system.

In order to compare a pair of systems we use the t-statistics for unequal variances to test whether the difference in the mean F_1 measure attained by these systems is statistically significant. Comparing the first and the second system, we find that the statistic admits the value 4.48833 with a probability of accepting the null hypothesis (i.e., the two means are equal) 0.0015 that is clearly below 0.05. Accordingly, the performance difference among these two systems is statistically significant for 0.05 level of significance. Similarly, when comparing the second and the third systems as well as the first and third systems the performance differences are found to be statistically significant for 0.05 level of significance.

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

[3] A. Tritschler and R. Gopinath, "Improved speaker segmentation and segments clustering using the bayesian information criterion," in Proc. *6th European Conf. Speech Communication and Techology*, pp. 679-682, September 1999.

[4] L. Lu and H. Zhang, "Speaker change detection and tracking in real-time news broadcast analysis," in *2004 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, pp. 741-744, June 2004.

[5] T. Wu, L. Lu, K. Chen, and H. Zhang, "UBM-Based Real-Time Segmentation for Broadcasting News," in Proc. *2003 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, pp. 193-196, April, 2003.

[6] 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, March, 2005.

[7] S. Cheng and H. Wang , "METRIC-SEQDAC: A hybrid approach for audio segmentation," in Proc. *6th Int. Conf. Spoken Language Processing*, October 2004.

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

[9] M. Kotti, E. Benetos, C. Kotropoulos, and L. G. P. M. Martins "Speaker change detection using BIC: A comparison on two datasets," in Proc. *2006 IEEE Int. Symp. Communications, Control, and Signal Processing*, March 2006.

[10] Y. Zhu and X. Rong, "Unified Fusion Rules for Multisensor multihypothesis network decision systems," in Proc. *2003 IEEE Trans. Systems, Man, and Cybernetics*, vol. 33, no 4, pp. 502-513, July 2003.

[11] F. van der Heijden, R. P. W. Duin, D. de Ridder, and D. M. J. Tax, *Classification, Parameter Estimation and State Estimation*, London UK: Wiley, 2004.

[12] J. P. Campbell, JR, "Speaker recognition: A tutorial", *Proccedings of the IEEE*, vol. 85, no. 9, pp.1437-1462, 1997.

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