Speaker Change Detection using BIC: A comparison on two datasets.

Margarita Kotti, Emmanouil Benetos,
Constantine Kotropoulos
Dept. of Informatics, Aristotle University of Thessaloniki, Greece
{mkotti,empeneto,costas}@aiia.csd.auth.gr

Luís Gustavo P. M. Martins
INESC Porto, Porto
Portugal
{lmartins}@inescporto.pt

Abstract— **This paper addresses the problem of unsupervised speaker change detection. We assume that there is no prior knowledge of the number of speakers or their identities. Two methods are tested. The first method uses the Bayesian Information Criterion (BIC), investigates the AudioSpectrumCentroid and AudioWaveformEnvelope features, and implements a dynamic thresholding followed by a fusion scheme. The second method is a real-time one that uses a metric-based approach employing line spectral pairs (LSP) and the BIC criterion to validate a potential speaker change point. The methods are tested on two different datasets. The first set was created by concatenating speakers from the TIMIT database and is referred to as the TIMIT data set. The second set was created by using recordings from the MPEG-7 test set CD1 and broadcast news and is referred to as the INESC dataset.**

I. Introduction

Speaker segmentation aims at finding the speaker change points in an audio stream. It is of extreme importance, as it is a preprocessing task for audio indexing, speaker identification - verification - tracking, automatic transcription, etc. Massive research has been carried out during the last decade in this area. Tritschler and Gopinnath proposed the use of the Bayesian Information Criterion (BIC) over the mel-cepstrum coefficients (MFCCs) [4]. Delacourt and Wellekens proposed a two-pass segmentation technique called DISTBIC [2]. Ajmera et al introduced an alternative of BIC, which does not need tuning, and some heuristics [3]. Meanwhile, novel features like the smoothed zero crossing rate (SZCR), perceptual minimum variance distortionless response (PMVDR), and the filterbank log coefficients (FBLC) were introduced by Huang and Hansen [10]. Another method is the so-called METRIC SEQDAC [9]. Finally, a hybrid algorithm is proposed, which combines metric-based segmentation with the BIC criterion and model-based segmentation with Hidden Markov Models (HMMs) [7].

The major contribution of this paper is in the comparative performance of two speaker segmentation systems. Both systems are based on the BIC and their efficiency is tested on two different datasets using the same experimental protocol. The first system utilizes the AudioSpectrumCentroid and AudioWaveformEnvelope [1], a dynamic thresholding which adapts its value according to the audio stream, and a fusion scheme which combines the partial results so as to achieve a better performance than that obtained by the same algorithm without fusion. A multiple pass algorithm is developed that employs a distinct feature in each pass. Each pass is executed independently from others. The second system is a real-time speaker change detection system. It is able to recognize speaker turn points with the shortest possible delay without having access to the whole speech stream. This scenario does not utilize iterative techniques, as the first system does, and imposes limitations on the computational load of the algorithm. Similar arguments were also explored in [5], [6], [8]. In the second system, the processing is divided into two main stages: In the first stage, a metric-based approach is implemented for coarse speaker segmentation using Line Spectral Pairs (LSP). In the second stage, the BIC criterion is used to validate the potential speaker change points detected previously.

The rest of this paper is organized as follows. In Section II, the two systems are described. Experimental results are demonstrated in Section III, and conclusions are drawn in Section IV.

II. Two Systems for Speaker Change Detection

A. The first system

A BIC-type criterion is applied [2], [3], [4], [7], [9] and the BIC variant proposed in [3] is used. Speaker change detection is formulated as a hypothesis testing problem. We assume that there are two neighboring chunks X and Y around time c_j and the problem is to decide whether or not a speaker change point exists on c_j. Let $Z = X \cup Y$.

Under H_0 there is no speaker change at time c_j. The maximum likelihood (ML) principle is used to estimate the parameters of the chunk Z that is modelled by a GMM of two components. Let us denote the GMM parameters estimated by the expectation-maximization (EM) algorithm as θ_z. The log likelihood L_0 is calculated as

$$L_0 = \sum_{i=1}^{N_x} \log p(x_i|\theta_z) + \sum_{i=1}^{N_y} \log p(y_i|\theta_z) \qquad (1)$$

where N_x and N_y are the total number of samples in chunks X and Y, respectively.

Under H_1 there is a speaker change at time c_j. The chunks X and Y are modelled by multivariate Gaussian densities whose parameters are denoted by θ_x and θ_y. Then, the log likelihood L_1 is given by

$$L_1 = \sum_{i=1}^{N_x} \log p(x_i|\theta_x) + \sum_{i=1}^{N_y} \log p(y_i|\theta_y). \qquad (2)$$

The dissimilarity is estimated by

$$d = L_1 - L_0 - \frac{\lambda}{2} \cdot \Delta K \cdot \log(N_x + N_y) \qquad (3)$$

where λ is the penalty factor (ideally 1.0) tuned according to data and ΔK is the number of the model parameters [2], [3]. If $d > 0$ then a local maximum is found and time c_j is considered to be a speaker change point. Otherwise, there is no change point at time c_j.

The selection of the appropriate features is of great importance since the accurate description of the audio signal is vital. We utilize the *mel cepstrum coefficients (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 of 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 of DFT magnitude, and in the last pass AudioWaveformEnvelope. The reason why multiple passes are employed is that after each pass, the number of chunks is decreased, because specific potential change points are discarded being false. So the length of chunks is becoming larger. Several researchers have come to the conclusion that the larger the chunks are, the better the performance is, because enough data are available for accurate speaker model estimation [2], [4], [5], [6], [8], [10]. 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. So for every speaker we keep the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ that 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 [5], [6], [8], [10].

The dynamic thresholding refers only to scalar features. We start with an ad hoc threshold t which may vary. The ad hoc threshold is determined after a considerable number of experiments during which we measure the efficiency of different thresholds and then we retain the threshold which maximizes the efficiency. 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 ϵ:

$$\epsilon = |f(k+1) - f(k)|. \tag{4}$$

Let $\bar{\epsilon}$ be the mean value of ϵ over all chunks of a recording. The value of $\bar{\epsilon}$ is compared to t, whose value is adjusted by adding or reducing 0.5% of $\bar{\epsilon}$ and the new adjusted threshold t' is:

$$t' = \begin{cases} t + 0.005 \cdot \bar{\epsilon} & \text{when} \quad t < \bar{\epsilon} \\ t - 0.005 \cdot \bar{\epsilon} & \text{when} \quad t > \bar{\epsilon}. \end{cases} \tag{5}$$

In order to estimate the GMM that is needed in (1) the EM algorithm is used, which may converge at local minima. However, there is no guarantee that this local minimum coincides with the global minimum or that there is only one local minimum. 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 introduced by the EM algorithm by repeating the experiment multiple times, say R times. In other words, a majority voting takes place in each pass. To be more specific, for each pass we obtain a set of possible speaker turn points. Let us denote it by $\mathcal{C}_i = \{c_1, c_2, \ldots, c_j\}$, where i is the running number of the experiment and c_1, c_2, …, c_j are the potential speaker change points. The final set of change points $\mathcal{C}_f$ of that pass consists of those potential speaker change points c_j 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 algorithm is summarized as follows.

```
1) Initialize $R, S, \mathcal{C}_f$=$\emptyset$
2) for $i = 1 : R$ find $\mathcal{C}_i = \{c_1, c_2, ..., c_j\}$
3) $\forall$ $c_j$ in all $\mathcal{C}_i$
   if Total-Count-Of-$c_j > S$  then $\mathcal{C}_f$=$\mathcal{C}_f \cup \{c_j\}$.
```

where by the term Total-Count-Of-c_j we denote the number of times the potential speaker change point c_j appears in all possible speaker turn points sets $\mathcal{C}_i$, $i = 1, 2, \ldots, R$. Diagrammatically, the algorithm is depicted as a directed graph in Figure 1 which represents a causal network in the horizontal direction. Apparently, pass1 affects pass2-pass10 and so on. Finally, f, f', f'' are the features that are implemented in each pass.

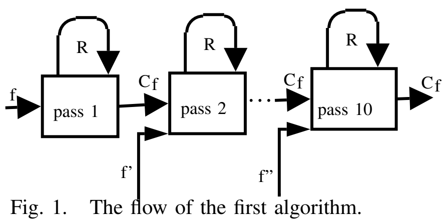

Fig. 1. The flow of the first algorithm.

B. The second system

The second system starts by down-sampling the input speech audio to 8 kHz, 16 bits mono channel format and applying pre-emphasis. The speech stream is then divided into frames with a duration of 25ms that do not exhibit overlap between each other. Zero crossing rate (ZCR), STE, and 10-order LSP features are extracted to a feature vector. The system considers only voiced parts of the speech signal, using ZCR and STE features to discard unvoiced and silence frames. In a first stage, speaker change detection is coarsely performed using a metric-based approach to calculate the distance between consecutive speech segments. Assuming that the LSP features are Gaussian distributed, each speech segment can be represented by a multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The Kullback-Leibler (K-L) divergence shape distance [11] is then used to estimate the distance between two speech segments i and j:

$$D(i,j) = \frac{1}{2}\mathrm{tr}[(\boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_j)(\boldsymbol{\Sigma}_i^{-1} - \boldsymbol{\Sigma}_j^{-1})] \tag{6}$$

where tr stands for the trace operator. The speech segments are formed by accumulating the necessary number of voiced frames until there are sufficient data to prevent ill-conditioned covariance matrices of the 10th-order LSPs used to model each speech segment. This implies that each speech segment should at least include 55 voiced frames which corresponds to a minimum segment duration of 1.375 sec of voiced speech. The hop size of the sliding of the segment window is 0.5 sec of voiced speech. As a consequence of this dynamic segment size and hop values, the absolute duration and start time of each speech segment is dynamically adjusted depending on the voiced content in each speech segment. This approach enables the system to be adapted to different speaking patterns and languages. On the other hand, the system output delay is not constant anymore, becoming dependent on this dynamic windowing design, which could be seen as a drawback. Using the presented K-L divergence shape distance, a potential speaker turn point can be detected between two segments whenever the following conditions are verified:

$$D(i, i+1) > D(i+1, i+2) \tag{7}$$

$$D(i, i+1) > D(i-1, i) \tag{8}$$

$$D(i, i+1) > Th_i. \tag{9}$$

The first two conditions guarantee that a local maximum exists. The third condition assures that a prominent distance peak is considered relevant. However, it is based on a threshold, whose value cannot be set trivially. Lu and Zhang [8] proposed an automatic threshold value defined as:

$$Th_i = \alpha \frac{1}{N} \sum_{n=0}^{N} D(i-n-1, i-n) \tag{10}$$

where N is the number of past distances considered, and α is a scaling factor. In this system we set N=3 and use α to tune the system response. Other techniques may be used to increase the confidence on the speaker change points identified by such a metric-based approach. One possibility is to use BIC to reduce the false alarm rate. Assuming two Gaussian models derived from two speech segments to be $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, the corresponding number of feature vectors as N_1 and N_2, and a single Gaussian model estimated using all feature vectors $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the BIC difference between the two models can be defined as:

$$BIC(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) = \frac{1}{2} \cdot \big((N_1 + N_2) \log |\boldsymbol{\Sigma}| - N_1 \log |\boldsymbol{\Sigma}_1| \\ - N_2 \log |\boldsymbol{\Sigma}_2|\big) - \frac{1}{2} \cdot \lambda(\delta + \frac{1}{2}\delta(\delta+1)) \log(N_1 + N_2) \quad (11)$$

where λ is the penalty factor for the model complexity, and δ is the feature dimension. If $BIC(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ takes 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.

Although the technique seems to be threshold free, in practice λ serves as a threshold that has to be fine-tuned manually. Furthermore, BIC demands high amounts of data in order to produce accurate results. Consequently, the system should not rely solely on the small amount of data available in the two consecutive speech segments. So as new speech data are received, the system should incrementally update the speaker's model, as happens in the first system. Though, in the second system we utilize an approach similar to the one proposed in [8]. Speaker models are stored using a quasi-GMM approach, a non-iterative solution that allows real-time operation. We propose a somewhat different implementation of the quasi-GMM procedure. Assuming that no speaker change is detected at a determined point, instead of discarding the arriving speaker data when the model reaches a number of 32 Gaussian mixtures, this implementation marks the oldest mixture in the current speaker model to be replaced by the new mixture created (or updated) from the new speech segment data.

As soon as this new mixture weight becomes significant (i.e. as it models an increasing number of speech frames, achieving a weight comparable to the one of the mixture marked to be replaced), the replacement operation is carried out. This mechanism is potentially more robust to the effects of speakers whose voices start to present changes after talking for long periods of time, or to long-term changes in background noises or recording conditions.

We are in fact comparing the current speech segment modelled by the multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, with the current quasi-GMM speaker model having S Gaussian densities denoted by $\mathcal{N}(\boldsymbol{\mu}_{1j}, \boldsymbol{\Sigma}_{1j})$, $j = 1, 2, \ldots S$, over N_{1j} feature vectors. The distance between the density model $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ can be roughly estimated as:

$$D = \sum_{j=1}^{S} w_{1j} \; BIC(\boldsymbol{\Sigma}_{1j}, \boldsymbol{\Sigma}_2) \quad (12)$$

where $w_{1j} = \frac{N_{1j}}{N_1}$ and $N_1 = \sum_{j=1}^{S} N_{1j}$. Whenever $D > 0$, the potential speaker change previously detected by the metric-based approach is confirmed as a real speaker boundary by the BIC refinement procedure.

III. Experimental Results

In order to assess the performance of the aforementioned algorithms two different datasets were utilized. The first set was created by concatenating speakers from the TIMIT database and is referred to as the TIMIT data set. 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. The second set was created by using recordings from the MPEG-7 test set CD1 and broadcast news and is referred to as the INESC dataset. In the second set, the audio format is PCM, the audio samples are quantized in 16 bit, the recordings are single-channel the mean duration is 19.81 sec and the st. dev. is 27.08 sec.

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:

$$FAR = \tfrac{FA}{GT + \, FA} \quad MDR = \tfrac{MD}{GT} \quad (13)$$

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:

$$PRC = \frac{CFC}{DET} \quad (14)$$

$$RCL = \frac{CFC}{GT} \quad (15)$$

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

$$F_1 = \frac{2.0 \cdot PRC \cdot RCL}{PRC + RCL} \quad (16)$$

that admits a value between 0 and 1. The higher its value is, the better performance is obtained. Between pairs FAR, MDR and PRC, RCL the following relationships hold:

$$MDR = 1 - RCL \quad (17)$$

$$FAR = \frac{RCL \cdot FA}{DET \cdot PRC + RCL \cdot FA}. \quad (18)$$

Tables I and II demonstrate the performance of the first system. In Table I the results for 10 randomly selected test recordings extracted from TIMIT database not included in the training set are demonstrated. The efficiency has been presumed dropping whenever the speaker's utterance has a duration of less than 1-2 sec, as it was expected [2], [4], [5], [6], [8], [10]. In Table II the respective results for 14 randomly selected files from INESC database are included.

Tables III and IV describe the performance of the second system. In Table III the results for the same 10 randomly selected test recordings extracted from the TIMIT database are demonstrated. In Table IV the respective results for 14 files from INESC database, which are the same as above, are shown. For both cases a is equal to 0.4 and λ is equal to 0.9.

IV. Conclusions

The performance of two BIC-based speaker segmentation systems were compared in this paper. The first system puts a higher emphasis on the accuracy than the real-time operation. The second system favors real-time operation at the expense of performance accuracy. Each system was evaluated on two datasets. In the first dataset, where short dialogues are present, the first system yields

TABLE I
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	0.00	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	0.78	0.70	0.72	0.218	0.305
st. dev.	0.137	0.136	0.008	0.135	0.136

TABLE II
PERFORMANCE OF THE FIRST SYSTEM ON THE INESC DATASET.

Index	PRC	RCL	F_1	FAR	MDR
1	0.12	0.78	0.20	0.88	0.22
2	0.30	0.69	0.42	0.70	0.31
3	0.36	0.77	0.49	0.64	0.23
4	0.14	0.71	0.23	0.86	0.29
5	0.08	0.56	0.14	0.92	0.44
6	0.13	0.67	0.22	0.87	0.33
7	0.12	0.78	0.21	0.88	0.22
8	0.03	0.25	0.05	0.97	0.75
9	0.12	1.00	0.22	0.88	0.0
10	0.19	0.58	0.29	0.81	0.42
11	0.13	0.58	0.21	0.87	0.42
12	0.45	0.72	0.56	0.55	0.29
13	0.03	0.17	0.05	0.97	0.83
14	0.19	0.58	0.29	0.81	0.42
mean	0.170	0.631	0.256	0.822	0.369
st. dev.	0.121	0.213	0.148	0.115	0.212

TABLE III
PERFORMANCE OF THE SECOND SYSTEM ON TIMIT DATASET.

Index	PRC	RCL	F_1	FAR	MDR
1	0.50	0.10	0.17	0.09	0.80
2	0.50	0.18	0.27	0.15	0.73
3	0.50	0.25	0.33	0.20	0.67
4	0.50	0.30	0.36	0.23	0.60
5	0.40	0.20	0.27	0.23	0.70
6	0.57	0.29	0.38	0.18	0.64
7	0.67	0.14	0.24	0.67	0.79
8	0.80	0.44	0.57	0.10	0.50
9	0.80	0.22	0.35	0.53	0.72
10	0.86	0.30	0.44	0.48	0.65
mean	0.609	0.242	0.338	0.135	0.679
st. dev.	0.160	0.097	0.115	0.072	0.088

TABLE IV
PERFORMANCE OF THE SECOND SYSTEM ON THE INESC DATASET.

Index	PRC	RCL	F_1	FAR	MDR
1	0.18	0.70	0.28	0.77	0.20
2	0.43	0.86	0.57	0.53	0.07
3	0.27	0.50	0.35	0.58	0.42
4	0.19	0.88	0.31	0.79	0.00
5	0.13	0.60	0.21	0.80	0.30
6	0.14	0.60	0.23	0.78	0.30
7	0.15	0.60	0.24	0.78	0.30
8	0.13	0.80	0.22	0.84	0.00
9	0.13	0.80	0.23	0.83	0.00
10	0.03	0.88	0.05	0.97	0.00
11	0.21	0.77	0.33	0.75	0.15
12	0.15	0.54	0.24	0.75	0.38
13	0.55	0.80	0.65	0.40	0.13
14	0.13	0.71	0.21	0.83	0.14
mean	0.200	0.712	0.294	0.743	0.172
st. dev.	0.134	0.128	0.151	0.146	0.149

an F_1 measure equal to 0.72 and outperforms the real-time system achieved an F_1 measure of 0.338. In the second dataset, where long dialogues are included, the second system attained an F_1 measure equal to 0.294. The first system achieved an F_1 measure of 0.256, because the mean duration of a speaker's utterance in INESC dataset (19.81 sec) is much longer than the mean duration in TIMIT dataset (3.28 sec) that the system was designed for. This leads to over-segmentation as can be seen from the large FAR.

ACKNOWLEDGEMENT

The joint work presented was developed within VISNET, a European Network of Excellence (http://www.visnet-noe.org), funded under the European Commission IST FP6 programme.

REFERENCES

[1] ISO/IEC 15938-4:2001, "Multimedia Content Description Interface - Part 4: Audio", Version 1.0.

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

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

[4] 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.

[5] 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.

[6] 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, Hong Kong, April, 2003.

[7] 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, March, 2005.

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

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

[10] 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.

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