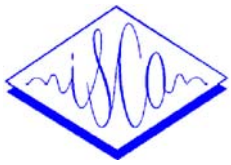

ODYSSEY04 - The Speaker and Language Recognition Workshop
Toledo, Spain
May 31 - June 3, 2004

UNSUPERVISED SPEAKER SEGMENTATION OF BROADCAST NEWS USING MDL-BASED GAUSSIAN MODEL

Jia-Hsin Hsieh and Chung-Hsien Wu

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan, R.O.C.
{ ngsnail, chwu }@csie.ncku.edu.tw

ABSTRACT

This paper proposes an approach for unsupervised speaker segmentation and gender discrimination of broadcast news. In this paradigm, a speaker segmentation mechanism using MDL-based Gaussian model is firstly adopted to determine the speaker changes using mean and covariance of the Gaussian model. These speaker segments partitioned by speaker changes are smoothed and discriminated into male or female. Experimental results show the proposed method achieved a better performance with 9.2% missed detection rate and 7.5% false alarm rate compared to the Delta-BIC method for speaker segmentation on broadcast news. In addition, the segment-based gender discrimination improves 9% accuracy compared to the clip-based discriminator.

1. INTRODUCTION

Automatic speaker segmentation and gender discrimination is very important for broadcast news transcription and indexing. Speech stream in broadcast news without proper speaker segmentation and gender discrimination will be very difficult for speech transcription. As these speech streams have been segmented into speaker segments, we can then recognize the speakers or genders and apply speech recognition technology to transcribe the content of the speech precisely [1]. There are several previous approaches on speaker change detection. Lu [2] applied threshold-based classifier for clip-based hierarchical audio classification and developed a linear spectral pair vector quantization approach with Bayesian Information Criterion (BIC) and adaptable GMM for speaker change detection. Mori [3] address the problem of speaker change detection and speaker clustering using VQ distortion. Adami [4] proposed a new approach to deal with the speaker segmentation of speech conversation, but the number of speakers was limited to two. Furthermore, Delta-BIC [5] is the most widely used approach for speaker change detection.

There are two major problems in previous speaker segmentation and gender discrimination tasks. Firstly, the Delta-BIC approach and others just detects one speaker change in an analysis window, and the small size of sliding window will degrade the segmentation performance and cause lots of false alarm. Secondly, the unit for clip-based speaker/gender recognition does not provide enough and meaningful information. The performance may degrade because of the unsuitable unit for speaker/gender recognition.

In order to deal with these problems, we propose a framework for speaker change detection and gender discrimination. In this approach, multiple speaker changes in mean vector and covariance matrix of an analysis frame based on minimum description length (MDL) criterion [6] are considered. Because we can detect multiple speaker changes in one analysis window, the effect of window size become insignificant. We do not need any prior information such as who the speaker is or how many speakers there are. Furthermore, because the classification unit is speaker segment, the segment-based gender discrimination approach is suitable.

2. PROPOSED SCHEME

In this paper, we propose an automatic speaker segmentation scheme which can determine multiple speaker changes in an analysis window of broadcast news. Generally, regardless of who the speakers are, we suppose the samples in different speaker segments have sharp changes of mean vector and covariance matrix in the feature space. We use an independently and identical (i.i.d.) multivariate Gaussian model based on MDL criterion to model each speaker segment. Given a general speech stream, feature analysis is conducted using a 16 ms frame shifted by 8 ms. For an analysis frame the Mel-Frequency Cepstral Coefficients (MFCCs) are extracted as the features and used to estimate the mean vector and covariance matrix of the Gaussian model. The proposed scheme is illustrated in Fig. 1. In this scheme, first, a heuristic silence deletion procedure identifies all silence frames in the speech stream. The speaker segmentation procedure detects the speaker changes and partitions the speech stream into several speaker segments with unknown speakers. Then the speaker segments are discriminated into male/female gender type. Finally a heuristic smoothing mechanism is used to remove short segments.

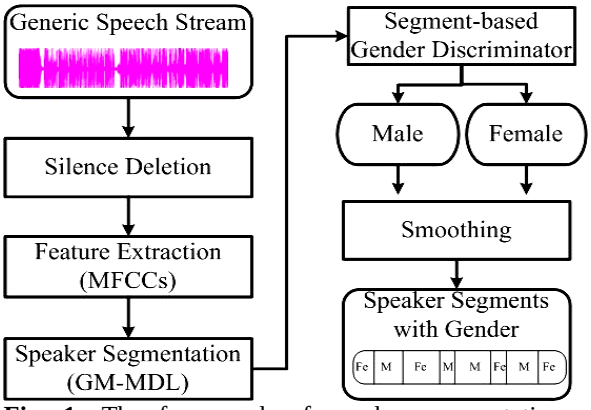

Fig. 1: The framework of speaker segmentation and gender discrimination

2.1. Speaker Segment Modeling

In speaker segment modeling using Gaussian model, the speech signal is transformed into a feature vector sequence $\mathrm{y}=(y_1, y_2, ..., y_n)$. Suppose **y** is a sequence of independent *d*-dimensional Gaussian random vectors with parameters $(\mu_1,\Sigma_1),(\mu_2,\Sigma_2),\ldots,(\mu_n,\Sigma_n)$, where μ_j and Σ_j is the mean vector and covariance matrix of the j^{th} segment, respectively. To determine the best speaker change positions in the sequence **y**, we introduce the hypothesis testing mechanism and test the following two hypotheses:

$$H_0: \mu_1=\mu_2=...=\mu_n \text{ and } \Sigma_1=\Sigma_2=...=\Sigma_n (\text{all } \mu,\Sigma \text{ are unknown})$$

versus the alternative

$$H_b: \mu_1=...=\mu_{c_1}\neq\mu_{c_1+1}=...=\mu_{c_2}\neq...=\mu_{c_b}\neq\mu_{c_b+1}=...=\mu_n$$
$$\text{and } \Sigma_1=...=\Sigma_{c_1}\neq\Sigma_{c_1+1}=...=\Sigma_{c_2}\neq....=\Sigma_{c_b}\neq\Sigma_{c_b+1}=...=\Sigma_n$$

where $\boldsymbol{H_0}$ represents there is no speaker change; $\boldsymbol{H_b}$ represents there are *b* speaker changes in the sequence **y** and the feature vector sequence **y** is divided into *b*+1 speaker segments; $\mathrm{c}=(c_1,c_2,...,c_b)$ is the sequence of the speaker change positions and *b* is the number of speaker changes. We suppose the feature vectors of each speaker segment y_i are drawn from an i.i.d. multivariate Gaussian distribution. The log likelihood function of **y** under hypothesis $\boldsymbol{H_b}$ is defined as:

$$\log f(\mathrm{y}\mid\mu,\Sigma,\mathrm{c},b)=-\frac{nd}{2}\log(2\pi)-\frac{1}{2}\sum_{j=1}^{b+1}m_j\log|\Sigma_j| - \frac{1}{2}\sum_{j=1}^{b+1}\sum_{i=c_{j-1}+1}^{c_j}(y_i-\mu_j)\Sigma_j^{-1}(y_i-\mu_j) \quad (1)$$

where m_j is the number of feature vector of the j^{th} speaker segment. c_0+1 denotes $\boldsymbol{y_1}$ and c_b+1 denotes $\boldsymbol{y_n}$.

Because all the parameters $\theta=(\mu_1,...,\mu_{b+1},\Sigma_1,...,\Sigma_{b+1})$ of **y** under hypothesis $\boldsymbol{H_b}$ are unknown, we employ the maximum likelihood estimator of θ to obtain $\hat{\theta}=(\hat{\mu}_1,...,\hat{\mu}_{b+1},\hat{\Sigma}_1,...,\hat{\Sigma}_{b+1})$ and maximize the log likelihood, where

$$\hat{\mu}_j=\bar{\mathrm{y}}_j=\frac{1}{m_j}\sum_{i=c_{j-1}+1}^{c_j}y_i,\ \hat{\Sigma}_j=\frac{1}{m_j}\sum_{i=c_{j-1}+1}^{c_j}(\mathrm{y}_i-\bar{\mathrm{y}}_j)(\mathrm{y}_i-\bar{\mathrm{y}}_j)' \quad (2)$$

Finally, we obtain the log likelihood for speaker segmentation of **y** under hypothesis $\boldsymbol{H_b}$:

$$\log f(\mathrm{y}\mid\hat{\theta},\mathrm{c},b)=-\frac{nd}{2}\log(2\pi)-\frac{1}{2}\sum_{j=1}^{b+1}m_j\log|\hat{\Sigma}_j|-\frac{nd}{2} \quad (3)$$

2.2. Speaker Change Detection

Minimum Description Length (MDL) [6] criterion was generally applied for model selection, hypothesis testing and so on. In this approach, we employ the MDL criterion for speaker segmentation. The MDL criterion is defined as

$$MDL(k)=-\log L(\hat{\Theta}_k)+\left\{\frac{k}{2}\log n+\left(\frac{k}{2}+1\right)\log(k+2)\right\} \quad (4)$$

In section 2.1, the $\log f(\mathrm{y}\mid\hat{\theta},\mathrm{c},b)$ represents the probability that there are *b* speaker changes inside **y**. In terms of the MDL criterion, three factors will affect the probability in speaker segmentation: the order of the model, the length of the data and the number of the speaker segments. We adopt the MDL criterion in our approach and obtain

$$MDL(\mathrm{y}\mid\hat{\theta},\mathrm{c},b)=-\log f(\mathrm{y}\mid\hat{\theta},\mathrm{c},b)+P(\mathrm{y},\mathrm{c},b) = -\left\{-\frac{nd}{2}\log(2\pi)-\frac{1}{2}\sum_{j=1}^{b+1}m_j\log|\hat{\Sigma}_j|-\frac{nd}{2}\right\}+\left\{(b+1)\times\left(\frac{k}{2}\log n+\left(\frac{k}{2}+1\right)\log(k+2)\right)\right\} \quad (5)$$

where $k=d+d(d+1)/2$ is the number of free parameters in multivariate Gaussian model (the order of the model). In Eq. (5), the lower score we obtain, the higher probability that there are *b* speaker changes inside **y**.

2.3. Hierarchical Binary Segmentation

In this section we develop a hierarchical binary segmentation procedure. We employ MDL criteria to decide when the segmentation procedure should stop to obtain the best solution $(\hat{\mathrm{c}},\hat{b})$. The basic idea to choose the speaker change is under the assumption that the suitable speaker change has lower MDL score. Given a sequence of *d*-dimensional Gaussian random vector $\mathrm{y}=(y_1,y_2,...,y_n)$, the procedure is described as follows:

Hierarchical Segmentation Procedure

Step1. Calculate $MDL(y\mid\hat{\theta},c,b=0)$

The score reveals the condition that there is no speaker change inside **y** ($b = 0$).

Step2. Obtain $\hat{c}^{b=1} = \arg\min_{d < c^{b=1} < n-d} MDL(y \mid \hat{\theta}, c^{b=1}, b = 1)$

Suppose there is one speaker change inside **y**. We calculate all the MDL scores by scanning all possible speaker change positions to obtain $\hat{c}^{b=1}$. The range of possible speaker change positions is between d and $n - d$ in order to obtain the maximum likelihood estimation of **y**.

Step3. Obtain $\hat{c}^{b=2} = \arg\min_{d < c^{b=2} < \hat{c}^{b=1} - d,\ \hat{c}^{b=1} + d < c^{b=2} < n-d} MDL(y \mid \hat{\theta}, (c^{b=2}, \hat{c}^{b=1}), b = 2)$

Suppose there are two speaker changes given the first speaker change $\hat{c}^{b=1}$. As Step2, we can obtain the second speaker change $\hat{c}^{b=2}$.

Step4. Repeat until $\frac{MDL(y \mid \hat{\theta}, (\hat{c}^{b=k+1}, ..., \hat{c}^{b=1}), b = k+1)}{MDL(y \mid \hat{\theta}, (\hat{c}^{b=k}, ..., \hat{c}^{b=1}), b = k)} > \lambda$

This inequality used to decide when the procedure should stop to obtain the final speaker change number and positions $(\hat{c}, \hat{b})$. λ is the convergence parameter determined by the desired sharpness of the speaker change and usually assigned to 1. If this inequality meets, we stop the procedure and obtain k speaker changes and their corresponding positions: $((\hat{c}^{b=k}, ..., \hat{c}^{b=1}), \hat{b} = k)$. Fig. 2 shows the speaker change detection results of the proposed approach with two speaker changes.

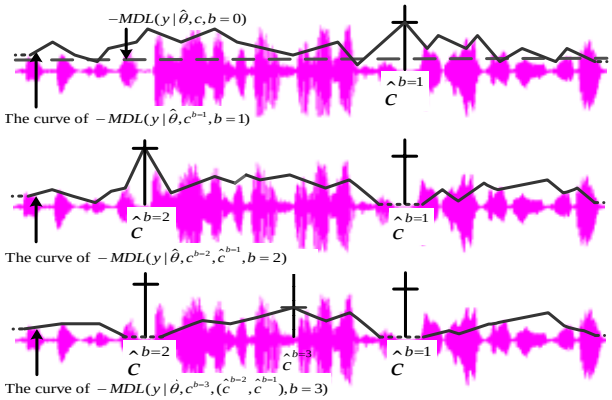

Figure 2: The speaker segmentation results of the proposed approach with two speaker changes.

2.4. Segment-based Gender Discrimination and Smoothing

Traditional clip-based gender discrimination may ignore small speaker change and result in discrimination errors because the clip length may be unsuitable for gender discrimination. This is because the samples inside the same clip may not be pronounced by the same speaker or gender. The segment-based gender discrimination could deal with the disadvantages of the clip-based approach because the classification unit is the segment with the same speaker or gender. Even though the speech stream may be segmented into many small sub-segments, the sub-segments still retain the homogeneous characteristic. A set of collected and pre-classified speech stream documents are used to construct a classification model. Here we employ the Gaussian Mixture Model (GMM) for gender discrimination. The smoothing procedure is a heuristic method which removes short fragments to decrease the false alarm probability for speaker segmentation.

3. EXPERIMENTS

3.1. Audio Corpus

In our experiments, we employed a part of Topic Detection and Tracking 3 (TDT-3) Mandarin audio corpus (1 hour long per file) which were recorded from Voice of America (VOA). The audio files in this corpus are mono and recorded with a sampling rate of 16 KHz. The statistics of the data are listed in Table 1.

Table 1: Part of the TDT-3 Mandarin audio corpus

Date_Time	*Number of Speaker change*
981001_08_09	107
981002_08_09	75
981003_08_09	102

3.2. Evaluation Measure

For performance evaluation of the segmentation task, the Detection Error Trade-off (DET) curve using Type-I errors: false alarm rate (FAR) and Type-II errors: missed detection rate (MDR) is suitable and is widely used in many previous approaches. Figure 3 shows an example of the MDR, FAR and change-point offset evaluation for the audio segmentation task. The tolerance is 3 sec for our segmentation experiments. The definition of MDR and FAR are

$$\text{MDR} = \frac{\text{number of missed change points}}{\text{number of target change points}} \times 100\% \quad (6)$$

$$\text{FAR} = \frac{\text{number of false alarm change points}}{\text{number of detected change points}} \times 100\%$$

These two evaluation scores can generate a DET curve after adjusting some system parameters.

In supervised audio classification, the training data are used to construct and then evaluate the audio classifier for inside testing. The testing data was further used to evaluate the performance of audio classifier for outside testing.

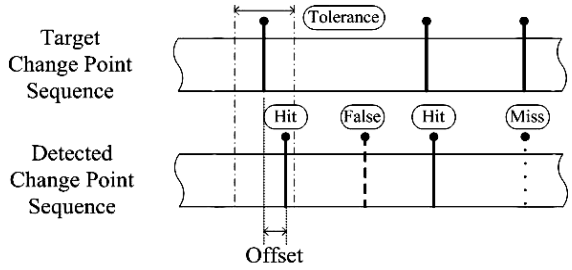

Figure 3: An example of MDR, FAR and offset of change-points

3.3. MDL-Based GM vs. Delta-BIC Approaches

Figure 4 compares the performance of MDL-based Gaussian model and Delta-BIC approach [5]. The MDL-based Gaussian model includes silence deletion and uses 26-dimensional MFCCs. The range of convergence parameter λ is from 0.8 to 1.1 with a step size of 0.05. We experimented on four different lengths of analysis window: 10sec, 20sec, 40sec and 60sec in our method. A convergence parameter λ of 0.95 and window length of 40 sec achieved the best performance with an MDR of 0.092 and an FAR of 0.075. The experimental results show the MDL-based Gaussian model outperformed the Delta-BIC approach. Furthermore, Table 2 shows the comparison of the average execution time for the MDL-based Gaussian model and Delta-BIC.

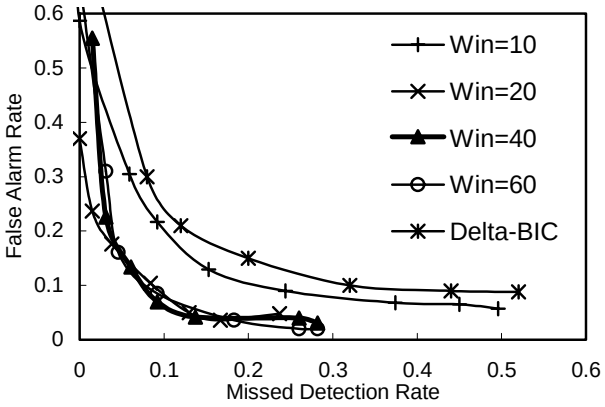

Figure 4: The DET curve of MDL-based Gaussian model vs. Delta-BIC approach

Table 2: The average execution time for 1 hour audio data

	Win=10	*Win=20*	*Win=40*	*Win=60*	*Delta_BIC*
Time (hr)	0.5	1	2	3	4

3.4. Performance of Gender Discrimination

For gender discrimination, the clip-based and segment-based GMM discriminators with 32 mixtures were constructed and evaluated. The feature for discriminator was 26-dimentional MFCCs. The results are shown in Table 3. From this table, the clip size in clip-based gender discriminator is the major problem and results in worse performance than the proposed segment-based approach.

Table 3: Comparison of the segment-based and clip-based gender discriminator

	Inside Testing	*Outside Testing*
Segment-based	0.92	0.87
Clip-based	0.85	0.78

4. CONCLUSION

This paper has presented an MDL-based Gaussian model for automatic speaker segmentation of broadcast news. The multiple speaker changes in mean vector and covariance matrix was used to segment the speech stream and the segment-based gender discriminator was employed to discriminate the speaker segment. The experimental results show that the proposed approach achieved better performance than Delta-BIC approach.

ACKNOWLEDGEMENT

The authors would like to thank the National Science Council, Republic of China, for its financial support of this work, under Contract No. NSC90-2213-E-006-088.

REFERENCE

[1] P. Beyerlein, X. Aubert, R. Haeb-Umbach, M. Harris, D. Klakow, A. Wendemuth, S. Molau, H. Ney and Michael Pitz, "Large vocabulary continuous speech recognition of Broadcast News – The Philips/RWTH approach," *Speech Communication*, vol. 37, pp. 109-131, 2002 .

[2] L. Lu, H.-J. Zhang ,and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Trans. on Speech and Audio Proc.*, vol. 10, pp. 504-516, 2002.

[3] K. Mori and S. Nakagawa, "Speaker change detection and speaker clustering using VQ distortion for Broadcast news speech recognition," *in Proc. of ICASSP'01*, pp. 413-416, 2001.

[4] Andre G. Adami, Sachin S. Kajarekar, Hynek Hermansky, "A new speaker change detection for two-speaker segmentation, " *in Proc. of ICASSP'02*, pp. 3908-3911, USA, 2002.

[5] S.S. Chen and P.S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion," in *Proc, of the DARPA Broadcast News TRanscri. & Underst. Workshop*, Landsdowne, VA, 1998.

[6] .J. Rissanen, "Stochastic complexity," *Journal of the Royal Statistical Society*, series B, vol. 49, pp. 223-239, 1987.