APSIPA ASC 2009, pp. 817-820, Oct. 4-7, 2009, Sapporo, Japan

Speaker segmentation based on between-window correlation over speakers' characteristics

Gang Wang and Thomas Fang Zheng
Center for Speech and Language Technologies, Division of Technical Innovation and Development,
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China
E-mail: Wang-g07@mails.tsinghua.edu.cn, Fzheng@tsinghua.edu.cn Tel/Fax: +86-010-62796393

***Abstract*—Speaker segmentation is widely applied in many domains such as multi-speaker detection and speaker tracking. However, the performance of the conventional metric-based methods is neither good enough nor stable due to the stability of the between-window distance calculation. In order to enhance the stability and hence to improve the performance, a new method based on the between-window correlation over speakers' characteristics is proposed. In this method, a set of reference speaker models are trained which can represent the whole speaker model space. The between-window correlation of likelihood vectors of scores against these reference models is taken as the metric. The gender information and the Peak and Valley information are also used. Experiments over NIST SRE 2002 Segmentation BNEWS and SWBD Datasets show that better performance can be achieved compared with the BIC and the GLR methods. What's more, the proposed method can achieve approximately the best performance in a wider value range of predefined thresholds than the BIC and the GLR methods, which reduces the threshold sensitivity.**

I. INTRODUCTION

The goal of speaker segmentation is to segment an utterance into acoustically homogeneous segments, each of which contains only one speaker. It has been widely applied in multi-speaker detection, speaker tracking and so on. The state-of-the-art methods can be classified into three categories: the metric-based, the model-based, and the hybrid of them.

In the metric-based segmentation, the distance between two adjacent analysis windows sliding over the utterance is calculated, and if it is greater than a predefined threshold, the boundary between the windows is regarded as a speaker change point. The commonly used distance measures include Bayesian Information Criterion (BIC) [1] 、Generalized Likelihood Ratio (GLR) [2]、Kullback-Leibler Divergence (KL) [3]、Support Vector Machine (SVM) [4], and Audio Entropy[5], and so on. In the model-based segmentation, it first of all estimates those possible target speakers from the utterance, then searches the target speakers' change points using corresponding speaker models which are updated iteratively in the whole process, traditional methods include GMM-based [6], Eigenvoice-based [7] and HMM-based [8] ones, etc. The hybrid method normally combines the former two methods together, for example ELISA [9] is a hybrid of the HMM-based method and the BIC method.

However the between-window distance calculation in the metric-based method is often inaccurate and not stable, or the speaker change points locating in the model-based method is biased, especially when the speaker change is possible very frequent. So we can observe that the performance of segmentation method seriously depends on the definition of thresholds for the metric-based method or the quality of initial target speaker models for the model-based method. In order to enhance the stability of the distance calculation for the metric-based method, a new method is proposed in this paper. A set of reference models are defined which can better represent the whole speaker model space. Instead of calculating the distance between two adjacent windows, we propose to calculate the between-window correlation of likelihood vectors, each of which consists of the likelihood scores of the feature of the corresponding (left or right) window against the reference models.

The definition of the reference models in this paper is different from that in the well known Anchor model [10], in which an exact speaker model can be found for the tested speaker. The speaker recognition task in this paper is assumed to be open-set speaker identification, so the above idea will be less useful. The definition of the reference models is in this way: based on the K-L distance measure, a sufficient number of speakers in the development set are classified into several classes using a kind of VQ technique such as K-means [11], and each class forms a reference speaker model which can represent the characteristics of the speakers in this class. This can be regarded as reference models definition based on speakers' characteristics.

This paper is organized as follows. In Section II, a detailed description about the speakers' characteristics and the segmentation algorithm are described. The experimental results are given in Section III. Conclusions are drawn in Section IV.

II. SPEAKERS' CHARACTERISTICS AND SEGMENTATION ALGORITHM

In this section, the metric definition based on speakers' characteristics and the proposed segmentation algorithm will be described in detail.

A. Speakers' Characteristics and Metric Definition

Speakers' characteristics means the representatives for speakers' model space. The model space described by a set of existing speaker models is clustered into several classes and each class is represented by a new speaker model called a reference speaker model, or a speaker's characteristics generated from all the speaker models in this class. Considering the gender of the speaker is more helpful in detecting speaker change points, the gender characteristics is used in terms of gender-dependent modeling.

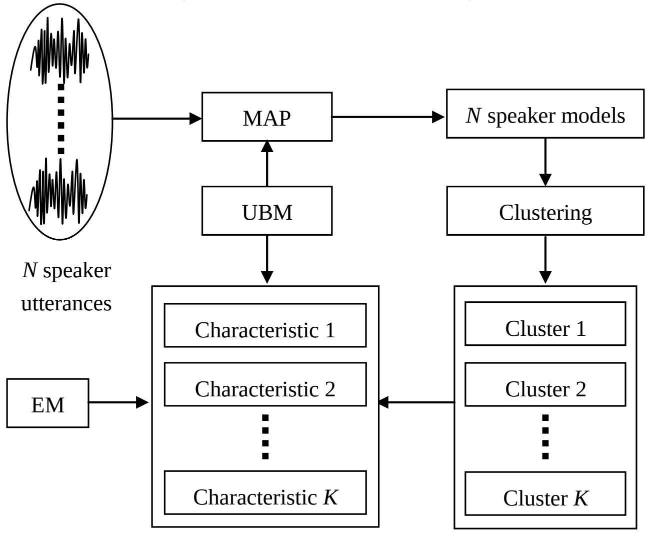

Fig. 1 Block Diagram of the Training of Speakers' Characteristics.

Here are the details. Let K denote the number of speakers' characteristics (reference speaker models) designed to cover the speaker model space. Speakers' characteristics are trained from the development set. In this dataset there are N utterances with each spoken by only speaker. The utterances are balanced over language, gender and speaking style, and the bigger the number of different speakers is the better the performance will be. N speaker models are then trained from the N utterances using GMM-UBM [12] modeling method and Maximum *a posteriori* estimation (MAP) [13]. N GMM supervectors will be obtained from the N speaker models by concatenating the mean vectors. The K-means [11] method is used to cluster the N supervectors into K classes. In each class, a quasi speaker model as its speakers' characteristics (reference model) is trained from the utterances in it based on the GMM-UBM [12].

Two gender-dependent UBMs are trained as speaker gender characteristics. The likelihood of an utterance against the speaker gender characteristics presents the gender information of speaker contained in this utterance.

Each speaker is modeled by a vector called Likelihood Vector, denoted by L_V, and defined as the likelihood scores of an utterance against those speakers' characteristics. Let C_i denote speakers' characteristics, where i is 1, 2, …, K, F, or M, and F and M are indices to the female and male characteristics, respectively. The Likelihood Vector is calculated as

$$L_v(X) = [P(X|C_i)]^T,\ i=1, 2, \ldots, K, F, \text{ or } M \qquad (1)$$

where X is an utterance, $P(X|C_i)$ is the log likelihood of X against characteristic C_i like in [12]. For two utterances X_1 and X_2, calculate $L_v(X_1)$ and $L_v(X_2)$ using Equation (1). The correlation coefficient between vectors $L_v(X_1)$ and $L_v(X_2)$ is defined as

$$\rho_{12} = C_{12}/\delta_1\delta_2 \qquad (2)$$

where C_{12} is covariance between $L_v(X_1)$ and $L_v(X_2)$, δ_1 and δ_2 are standard deviations of $L_v(X_1)$ and $L_v(X_2)$, respectively.

Just as in the VQ technique, the larger the number of speakers for training is, and the greater the value of K is, the more accurate the between-window correlation coefficient calculation will be. But the larger number of speakers and the greater value of K will lead to lower efficiency and training data sparseness. According to our experience, N=1,227 and K =300 can achieve a reasonable good performance.

B. Segmentation Algorithm

Based on the introduction of the speakers' characteristics as well as the metric, the segmentation is easier. It consists of following steps: (1) voice activity detection (VAD); (2) feature extraction; (3) speaker change point detection; (4) peak and valley validation. Steps (3) and (4) are illustrated in Fig.2 and Fig. 3, respectively, which are also to be described in details.

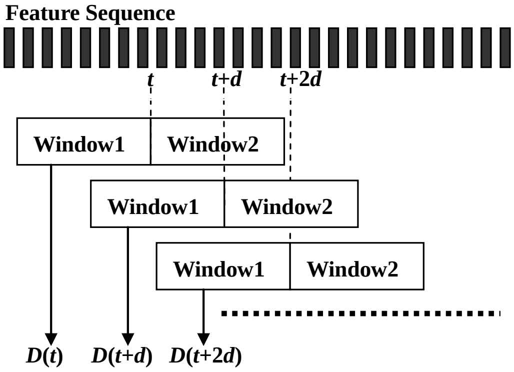

Fig. 2 Speaker Change Point Detection.

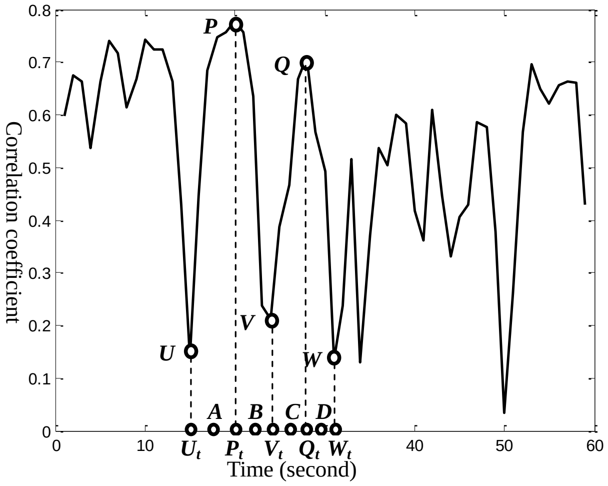

Fig. 3 Correlation Coefficient Curve.

Two same-sized adjacent sliding windows are used in Step (3) when calculating the correlation coefficients. The correlation coefficient between two neighboring windows at frame t, $D(t)$ is computed using Equations (1) and (2) as illustrated in Fig. 2. The window size of Windows 1 and 2 should meet the stability assumption of signal analysis and feature extraction. d, the window shift, is the difference of starting points of Window 1 between two adjacent calculation of the correlation and it stands for the resolution of segmentation algorithm. After the two windows sliding from left to right along the feature sequence, a curve of correlation coefficients can be obtained as in Fig. 3.

There are two assumptions.

H_0: if the speakers of two neighboring windows are the same, the correlation coefficient between the two windows' likelihood vectors is bigger.

H_1: if the speakers of two neighboring windows are different, the correlation coefficient between the two windows' likelihood vectors is smaller.

The assumptions indicate that the smaller the correlation coefficient is, the more likely there is a speaker change point at the boundary between Windows 1 and 2. However the problem is that how small the correlation coefficient should be when there is a true speaker change. Experimentally, we define a threshold for speaker change detection as follows. Say a putative speaker change at frame t is found if the following conditions are met

$$\begin{aligned} &|D(t) - D(t)_{lmax}| > \alpha\delta \\ &|D(t) - D(t)_{rmax}| > \alpha\delta \end{aligned} \qquad (3)$$

where $D(t)$ is a local minimum correlation coefficient value at frame t, and $D(t)_{lmax}$ and $D(t)_{rmax}$ are its left and right neighboring local maximum correlation coefficient values, δ is the standard deviation of correlation coefficient sequence, and α is an adjustable factor. Here $\alpha\delta$ can be regarded as the threshold for speaker change detection.

C. Peak and Valley Validation

The validation process is to determine if a putative speaker change point found in Step (3) is true or not. The process can be seen from the example given below. In Fig. 4, points *U, V, and W* are three adjacent valleys in the correlation coefficient curve, points *P* and *Q* are two peaks among them. Denote R_t as the corresponding horizontal coordinate (i.e. the time coordinate) of point *R* on the curve. Define:

$$\begin{aligned} &A = (U_t + P_t)/2,\ B = (P_t + V_t)/2 \\ &C = (V_t + Q_t)/2, D = (Q_t + W_t)/2 \end{aligned} \qquad (4)$$

Calculate the correlation coefficient of the utterance between *A* and *B* and the utterance between *C* and *D* as peak correlation D_p, the correlation coefficient of utterance between U_t and *A* and the utterance between *B* and V_t as valley correlation D_v. Actually, they can be regarded as to correspond to the most stable and the most unstable parts of the two segments, respectively.

If $D_p > \beta$, *V* is possibly not a speaker change point, delete *V* from the putative speaker change point set.

If $D_v < \gamma$, it is possibly that there should exist a speaker change point between U_t and V_t. We enlarge the resolution to analyze this part again using a window size and a window shift half of those in the previous analysis. Accordingly the putative speaker change point set will be changed or not. Here β and γ are two predefined thresholds experimentally learned from the development set. In our experiments, when $\beta = 0.4$ and $\gamma = 0.3$, best results can be achieved.

III. Experimental Results and Analysis

A. Experimental Data and Set up

The experiments were conducted based on GMM-UBM. The databases used for the speaker segmentation experiments were taken from National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE) [15] 2002 speaker segmentation data set. NIST SRE 2004 1C4W dataset was used to train the gender-dependent/independent UBMs through EM algorithm [14]. NIST SRE 2005, 2006, and 2008 8C4W datasets were used to train the speakers' characteristics. A broadcast news dataset named BNEWS and a telephone conversation dataset named SWBD were used as test sets. All the utterances were sampled at 8 kHz with 8-bit width.

Feature extraction were performed on a 20ms frame every 10ms. The pre-emphasis coefficient was 0.97 and hamming windowing was applied to each frame. An energy-based VAD was performed with each frame labeled either valid or invalid. 16-dimensional MFCC features were extracted from the utterances only for those valid frames with 30 triangular Mel filters used in the MFCC calculation. For each frame, the MFCC coefficients and their first derivative formed a 32-dimentional feature vector. The cepstral mean subtraction [16] in the feature-domain and session variability subspace projection [17] in the model-domain were applied to reduce the affect of channel. The UBM or each speaker characteristic was represented by an $M = 1{,}024$ Gaussian mixture density function, where the value of M was chosen empirically.

The window size is usually dependent on the speakers' change frequency of the data set. In our experiments the window size was chosen as 2s for BNEWS and 1s for SWBD according to performance experimentally while the window shift as 300ms for BNEWS and 100ms for SWBD according to resolution and efficient, respectively.

B. Experimental Results

False alarm rate (*FAR*) [15] and miss detection rate (*MDR*) [15] are used to evaluate the performance of the segmentation algorithm, which are defined as

$$\begin{aligned} FAR &= FA / (ASC + FA) \\ MDR &= MD / ASC \end{aligned} \qquad (5)$$

where *FA* denotes the number of false alarms, *MD* the number of miss detections, and *ASC* the actual number of speaker change points. If the time difference between a reference speaker change point and its nearest putative speaker change point is shorter than 300ms, this is regarded as a true detection, otherwise it is a miss. Results are given in tables I-IV.

In the following tables, the BIC and the GLR methods were chosen as the baseline systems. The proposed method based on Speakers' Characteristics is abbreviated as SC. G means the gender information was used while PV the Peak and Valley Validation used.

Comparison experiments show that the gender information is helpful for segmentation and the Peak and Valley Validation is useful to further reduce *FAR* and *MDR*. The stability of the distance calculation and accuracy of speakers' characteristics are enhanced compared with BIC and GLR, and SC outperformed BIC and GLR on both datasets though the improvement was not significant for SWBD dataset.

TABLE I
THE EFFECT OF THE NUMBER OF SPEAKERS' CHARACTERISTICS

K value	Test Dataset	*FAR*	*MDR*
100	BNEWS (SC) (Window Size = 2s)	35.7%	13.1%
200		34.2%	12.4%
300		32.8%	11.7%
400		32.9%	11.9%
100	SWBD (SC) (Window size = 1s)	45.3%	38.9%
200		44.1%	37.8%
300		42.7%	34.1%
400		42.8%	34.7%

TABLE II
EFFECT OF WINDOWS SIZE

Window Size (s)	Test Dataset	*FAR*	*MDR*
1.0	BNEWS (SC) (*K*=300)	35.7%	10.8%
1.5		34.2%	11.3%
2.0		32.8%	11.7%
2.5		31.9%	12.5%
0.8	SWBD (SC) (*K*=300)	45.3%	33.7%
1.0		42.7%	34.1%
1.5		42.2%	34.7%
2.0		41.6%	35.1%

TABLE III
PERFORMANCE COMPARISON WHEN G AND PV ARE USED OR NOT

Speakers' Characteristics	Test Dataset	*FAR*	*MDR*
SC	BNEWS (*K*=300) (Window size = 2s)	32.8%	11.7%
SC+G		33.1%	11.3%
SC+PV		32.2%	10.9%
SC+G+PV		31.5%	10.6%
SC	SWBD (*K*=300) (Window size = 1s)	42.7%	34.1%
SC+G		43.3%	33.8%
SC+PV		42.4%	33.6%
SC+G+PV		42.1%	33.2%

TABLE IV
COMPARED WITH BIC AND GLR

Methods	Test Dataset	*FAR*	*MDR*
BIC	BNEWS (*K*=300) (Window size = 2s)	33.8%	15.8%
GLR		34.2%	16.5%
SC+G+PV		31.5%	10.6%
BIC	SWBD (*K*=300) (Window size = 1s)	43.8%	35.1%
GLR		41.2%	34.3%
SC+G+PV		42.1%	33.2%

IV. CONCLUSIONS AND FUTURE WORK

In this paper we propose a speaker segmentation method based on between-window correlation over speakers' characteristics. Using the correlation of likelihood vectors of utterances against speakers' characteristics, the algorithm can remarkably improve the segmentation result in practice. Our experiments have shown that the speakers' characteristics and the gender information together are helpful for speaker segmentation. From the experiments we can find that the proposed method can achieve approximately the best performance in a wider value range of the adjustable factor than the BIC or the GLR method does. In other words, the performance of the proposed method depends on the definition of the threshold to a quite small extent.

Enough training data is needed to achieve a good performance for the proposed method, which is a shortcoming and needs further study.

REFERENCES

[1] S. S. Chen and P. S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion. in DARPA Speech Recognition Workshop, 1998

[2] H. Gish, M. H. Siu and R. Rohlicek. Segregation of speakers for speech recognition and speaker identification. In IEEE International Conference on Acoustics Speech and Signal Processing. 1991, 873-876

[3] M. A. Siegler, U. Jain, B. Raj and R. M. Stern. Automatic segmentation classification and clustering of broadcast news audio. in DARPA Speech Recognition Workshop, 1997, 97-99

[4] B. Fergani, M. Davy, A. Houacine. Speaker diarization using one-class support vector machines. Speech Communication 50 2008, 355–365

[5] J. M. Bai, S. W. Zhang, B. Xu. The technology of target speaker tracking in broadcast news. Acoustic Technology, 2005，(24):234-238

[6] M. C. Ivan, E. R. Aaron and S. Parthasarathy. Detection of target speakers in audio databases. ICASSP, 1999, 821-824

[7] C. Fabio, C. Daniele, et al. Stream-based speaker segmentation using speaker factors and Eigenvoices, ICASSP, 2008, 4133-4136

[8] S. Meignier, J. F. Bonastre and S. Igounet. E-HMM approach for learning and adapting sound models for speaker indexing. in 2001: A Speaker Odyssey, Chania, Crete, June 2001. 175-180

[9] D. Moraru, S. Meignier, C. Fredouille, et al. The ELISA consortium approaches in broadcast news speaker segmentation during the NIST 2003 rich transcription evaluation, ICASSP, 2004, 373-376

[10] M. Collet, D. Charlet and F. Bimbot. A correlation metric for speaker tracking using anchor models, ICASSP, 2005, 713-716

[11] A. V. Hall. Methods for demonstrating resemblance in taxonomy and ecology, Nature, Vol. 214, pp. 830-831, 1967

[12] D. A. Reynolds, T. Quatieri, R. Dunn. Speaker verification using adapted Gaussian Mixture Models [J]. Digital Signal Processing, 2000, 10: 19-41

[13] J. L. Gauvain, and C. H. Lee. Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains, IEEE Trans. Speech Audio Process, 2, 1994, 291–298

[14] A. Dempster, N. Laird and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. J. Roy. Stat. Soc. 1977, 39:1–38

[15] NIST Speaker Recognition Evaluation Plan, Online Available http://www.nist.gov/speech/tests/sre/

[16] S. Furui. Cepstral analysis technique for automatic speaker verification. IEEE Trans. on Acoustics, Speech and Signal Processing, 1981. 29(2):254-272

[17] J. Deng, T. F. Zheng, W. H. Wu. Session variability subspace projection based model compensation for speaker verification. ICASSP, 2007, IV， 57-60