UBM Based Speaker Segmentation and Clustering for 2-Speaker Detection

Jing Deng, Thomas Fang Zheng, and Wenhu Wu

Center for Speech Technology, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing, 100084
dengj02@mails.tsinghua.edu.cn, fzheng@tsinghua.edu.cn, wuwh@tsinghua.edu.cn

Abstract. In this paper, a speaker segmentation method based on log-likelihood ratio score (LLRS) over universal background model (UBM) and a speaker clustering method based on difference of log-likelihood scores between two speaker models are proposed. During the segmentation process, the LLRS between two adjacent speech segments over UBM is used as a distance measure, while during the clustering process, the difference of log-likelihood scores between two speaker models is used as a speaker classification criterion. A complete system for NIST 2002 2-speaker task is presented using the methods mentioned above. Experimental results on NIST 2002 Switchboard Cellular speaker segmentation corpus, 1-speaker evaluation corpus and 2-speaker evaluation corpus show the potentiality of the proposed algorithms.

Keywords: Speaker segmentation, Speaker clustering, Multi-speaker, Speaker Detection.

1 Introduction

In real-world speaker verification tasks over telephone, there is an increasing demand that speaker verification systems can verify one specific speaker whether in a conversation or not. One of the solutions to this demand is speaker segmentation and clustering. The aim of speaker segmentation and clustering is to segment an *N*-speakers' conversation into speech segments containing the voice of only one speaker (segmentation process) and to merge those speech segments belonging to a same speaker into one speech segment (clustering process). After speaker segmentation and clustering, a multi-speaker verification task can be simplified into several *N* single-speaker verification tasks. Generally, no *a priori* information is available on the number and identity of speakers involved in the conversation.

Previous researches have focused on two directions, distance based and model based. The former does not require any *a priori* information, but it is difficult to accurately describe the characteristics of a speaker with short speech segments which often occur in conversations over telephone and hence will result in a dissatisfactory performance during the clustering process. Methods in this direction include Bayesian Information Criterion (BIC) [1], [2], [3], Generalized Likelihood Ratio (GLR) [4], [5], Kullback-Leibler (KL) Distance [6], [7], DISTBIC [8], *etc*. The latter can achieve

Q. Huo et al. (Eds.): ISCSLP 2006, LNAI 4274, pp. 116 – 125, 2006.

a satisfactory result by building a model for each speaker in the audio recording and then using a global maximum likelihood score to find the best time-aligned speaker sequence (usually by using Viterbi algorithm). One of the difficulties in model based method is how to accurately build initial speaker models. The model based systems include LIA [9], ELISA [10], [11], *etc.*

Usually, there are many short speech segments in conversations over telephone. Distance based segmentation criteria, such as BIC, have some difficulties in dealing with them [8]. The reason is that it is difficult to estimate the characteristics of a speaker with short speech segment. Model based segmentation can well deal with this issue, however, they need *a priori* knowledge of speakers in the conversation. In order to well describe the characteristics of short speech segments, in this paper, UBM is used as *a priori* knowledge of speakers during segmentation process. Given two adjacent short speech segments belonging to a same speaker, the log-likelihood ratio score (LLRS) of them over UBM is small, and vice versa. So LLRS over UBM is used as a distance measure for speaker segmentation.

After segmentation, a conversation is divided into several speech segments. But the identity of each speech segment and the number of speakers are unknown. Because most conversations over telephone each contain only two speakers, the number of speakers in a conversation is set to 2 in this paper. Conventional speaker clustering methods mainly focus on finding out the closest speech segments while in this paper a method based on the difference of log-likelihood score between two speaker models is proposed to identify one speech segment as speaker *A* if it is the farthest one from speaker *B*. Over the NIST 2002 2-speaker segmentation Switchboard set, a system integrated with the proposed method can achieve a frame error rate of 6.8%, which will be detailed later.

This paper is organized as follows. The speaker segmentation based on LLRS will be presented in Section 2, and the proposed speaker clustering method will be described in Section 3. In Section 4, experiments and results will be described. Finally, conclusions and perspectives will be given in Section 5.

2 Speaker Segmentation Based on LLRS over UBM

In this paper, a simple segmentation criterion based on LLRS over UBM is used. First, acoustic features are extracted from the input speech. Then the acoustic features are divided into several decision windows by a sliding window with a 2-second width and a 0.1-second shift. In each decision window, the acoustic features are divided into two parts $X_1=(x_1, x_2, ..., x_i)$ and $X_2=(x_{i+1}, x_{i+2}, ..., x_N)$; and *LLRS* (i) between them is defined as

$$LLRS(i) = abs\left(L(X_1 \mid UBM) - L(X_2 \mid UBM)\right) \tag{1}$$

where i was set to the half position of the decision window. Because there may be some silence or noise in one decision window, the log-likelihood score of a speech frame over UBM is used as a measure to decide whether current frame is a speech frame or a non-speech frame. The bigger the log-likelihood score, the more likely current frame is a speech frame. A similar process is proposed in [12] which used the

log-likelihood score of one speech segment over UBM to separate the speech segment into three groups: *confidential* speech frames, *doubtable* speech frames, and *non*-speech frames. So in Equation (1), the acoustic features in each half decision window used to calculate the log-likelihood score are those whose scores are among the top half.

Finally, we can get a sequence of LLRS and the standard deviation σ can be estimated accordingly. In the LLRS plot (showed in Fig. 1), a peak is assumed to be a possible speaker turn point if

$$\left|max - min_l\right| > \alpha\sigma \quad \text{and} \quad \left|max - min_r\right| > \alpha\sigma \tag{2}$$

where α is an experiential value which is set to 0.5 in experiments in this paper, *max* is the LLRS at the peak position, and min_l and min_r are the left and right minima around the peak value point, respectively. More details about Equation (2) were described in [8].

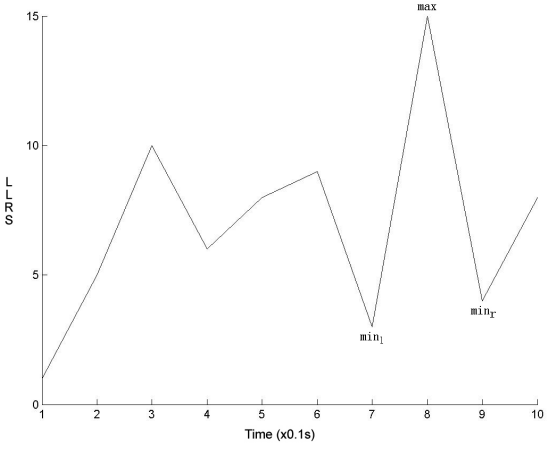

Fig. 1. LLRS plot: decision of a speaker turn

3 Speaker Clustering Based on Difference of Log-Likelihood Score Between Two Speaker Models

The goal described here in this section is to cluster speech segments with a same speaker identity. As mentioned above, the number of speakers in one conversation is 2. So given two speaker models (*A* and *B*) and several speech segments $\{X_i, i=1, 2, \ldots, N\}$, speech segment X_j is regarded to most likely belong to speaker model *A* if

$$j = \arg\max_i \left(L(X_i \mid A) - L(X_i \mid B) \right) \tag{3}$$

where *L*(.) is the log-likelihood function. After speaker segmentation, there are many short speech segments which are not long enough to well train a speaker model. In

order to solve this problem, a multi-stage clustering strategy is used. First, a UBM with a small number of components is used to select suitable speech segments for initial model training. Then with sufficiently long speech segments, speaker models can be well trained from a UBM with large number of components. The proposed speaker clustering method is described as follows.

Stage 1. Initial clustering

1.1 First an initial speaker model S_0 is adapted on the whole test utterance from *UBM*1 by MAP with only mean vector changed.

1.2 After speaker segmentation, all the speech segments are scored on S_0. The speech segment with the maximal log-likelihood score and longer than 2 seconds is selected for use of adapting speaker model S_1 from *UBM*1.

1.3 The remained speech segments are scored against S_0 and S_1, respectively. The difference of log-likelihood score, ΔS, is defined as

$$\Delta S = L(X \mid S_0) - L(X \mid S_1) \tag{4}$$

where X is the acoustic feature sequence from a speech segment. The bigger the ΔS is, the more likely X not belongs to S_1. The speech segment with the maximal ΔS and longer than 2 seconds is selected for use of adapting speaker model S_2 from *UBM*1.

1.4 Score the remained speech segments against S_1 and S_2. From those speech segments with score $L(X|S_1)$ bigger than $L(X|S_2)$, the speech segment with the maximal ΔS_{12} and longer than 1 second is selected for use of updating S_1, where ΔS_{12}= $L(X|S_1)$-$L(X|S_2)$. From those speech segments with score $L(X|S_2)$ bigger than $L(X|S_1)$, the speech segment with the maximal ΔS_{21} and longer than 1 second is selected for use of updating S_2, where ΔS_{21}= $L(X|S_2)$-$L(X|S_1)$.

1.5 Repeat 1.4 until there is no speech segment longer than 1 second.

1.6 Use S_1 and S_2 to calculate ΔS_{12} in speech segments belonging to S_1 and ΔS_{21} in speech segments belonging to S_2.

Stage 2. Refine the clustering

2.1 Adapting a new speaker model S_1 from *UBM*2 with speech segments belonging to previous S_1 which ΔS_{12} is among the top half.

2.2 Adapting a new speaker model S_2 from *UBM*2 with speech segments belonging to previous S_2 which ΔS_{21} is among the top half.

2.3 Score each speech segment against S_1 and S_2, respectively. If ΔS_{12} is positive, the speech segment is assigned to S_1, otherwise to S_2. Meanwhile, calculate ΔS_{12} on those speech segments belonging to S_1 and ΔS_{21} on speech segments belonging to S_2 for use in stage 3.

Here, *UBM*1 and *UBM*2 can be of different component sizes. In our experiments, *UBM*1 contains 16 components and *UBM*2 contains 1,024 components.

4 Experiments and Results

The features were extracted from speech signal at a frame size of 20 milliseconds every 10 milliseconds. The pre-emphasis factor was set to 0.97. The Hamming windowing was applied to each pre-emphasized frame. After that, a 256-point FFT was calculated for each frame and a bank of 30 triangular Mel filters were used.

Finally DCT was performed and 16-dimensional MFCC coefficients with the delta coefficients were obtained for each frame.

The baseline system in our experiments was based on the Gaussian Mixture Model-Universal Background Model (GMM-UBM) [13] with the UBM gender-independent, tree-structured [14], and containing 1,024 mixtures. No score normalization method was performed.

4.1 Segmentation Results

We tested the segmentation and the clustering methods on the NIST 2002 Switchboard Cellular speaker segmentation corpus. This corpus contains 199 test segments (two minutes each) involving only two speakers (at an 8 kHz sampling rate). The evaluation method was the NIST official scoring (version 07) [15] which is a frame based error rate protocol. Table 1 shows the accuracy of initial speech segments selection for model S_1 and S_2 in the clustering process (Steps 1.1 to 1.3). The segmentation results of LIA and the proposed method on NIST 2002 Switchboard Cellular speaker segmentation corpus are showed in Table 2.

The LIA system is an HMM based speaker segmentation system. Each state of the HMM characterizes a speaker and the transitions model the changes between speakers. During the segmentation, the HMM is generated using an iterative process, which detects and adds a new state (i.e. a new speaker) at each iteration.

We also compared the false alarm rates and the miss detection rates among BIC, GLR, DISTBIC, and the proposed method, listed in Table 3.

Table 1. Initial speech segments selection results on NIST 2002 Switchboard Cellular speaker segmentation corpus

Error Type	Error Time Rate (%)
Missed Speaker Time	0.1
False Alarm Speaker Time	0.3
Speaker Error Time	0.4

Table 2. NIST 2002 speaker segmentation results for Switchboard Cellular speaker segmentation corpus

System	Missed Speaker Time	False alarm Speaker Time	Speaker Error Time
LIA [16]	0.0%	0.0%	7.4%
Propose method	0.1%	0.1%	6.6%

Table 3. Segmentation performance comparison of BIC, GLR, DISTBIC, and the proposed method on the NIST 2002 Switchboard Cellular Speaker Segmentation Corpus

System	FAR(%)	MDR(%)
BIC	25.2	35.6
GLR	33.2	19.5
DISTBIC	30.8	20.3
Propose method	29.3	18.9

The total segmentation error rate of CLIPS is 8.6% and the fusion of LIA and CLIPS can achieve an error rate of 5.7% on NIST 2002 Switchboard Cellular speaker segmentation corpus [10]. Compared with LIA and CLIPS, the proposed method can achieve a comparative performance.

4.2 1-Speaker Detection (1D) Results

The training set contains 330 speech segments (two minutes each) by 139 males and 191 females. The test set contains 3,570 speech segments by 1,442 males and 2,128 females with about 15 to 45 seconds for each segment. The detection results are given in Fig. 2. Comparison result of LIA is showed in Fig 3 [16].

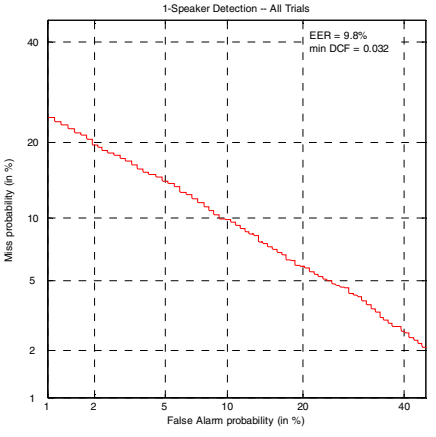

Fig. 2. 1-speaker detection results, NIST 2002 evaluation

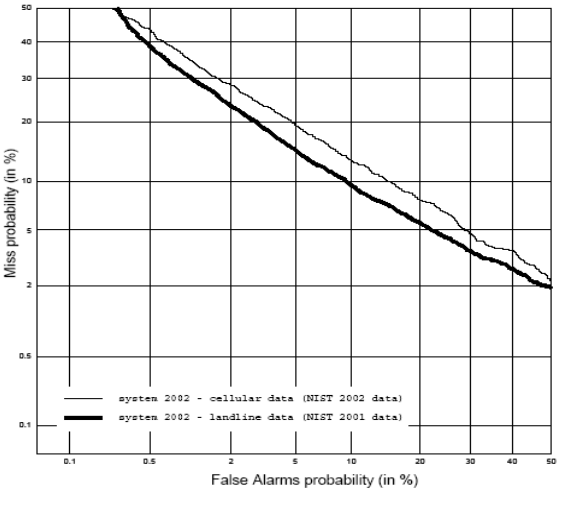

Fig. 3. LIA 1-speaker results on NIST 2002 cellular data and NIST 2001 landline data

4.3 1-Speaker Training, 2-Speaker Detection (1T- 2D) Results

Here, 1T-2D means using 1-Speaker speech segment for training and using 2-Speaker speech segment for detection. The training set here was same as that used in 1D evaluation. The test set contains 1,470 speech segments (one minute each) by 2 speakers (two males, two females or one male - one female). The detection results are given in Fig. 4.

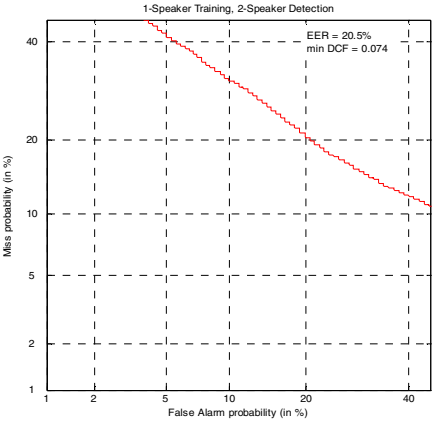

Fig. 4. 1T-2D results, NIST 2002 evaluation

4.4 2-Speaker Detection (2D) Results

This evaluation illustrates the effect of training a target speaker model from three 2-speaker audio files. No *a priori* information was provided except that the target speaker was the only speaker in each of the three files. The training set contains 309 target speakers (131 males and 178 females) and the test set contains 1,460 segments, each with an average duration of one minute spoken by two speakers. The training process is illustrated in Fig.5.

For each 2-speaker audio file, two final speech segments will be obtained by using the proposed segmentation and clustering methods. For each final speech segment, a speaker model can be trained form UBM with mean vectors changed only. That is to say, given three 2-speaker audio files, six speaker models can be obtained finally. Let S_1 and S_2 be any two speaker models form two audio files respectively, where the i-th components in S_1 and S_2 are defined as $\left(w_i, \mu_i^1, \Sigma_i\right)$ and $\left(w_i, \mu_i^2, \Sigma_i\right)$, respectively. The KL distance between S_1 and S_2 was calculated as

$$KL\left(S_1, S_2\right) = \sum_{m=1}^{M}\left(w_i \cdot \left(\mu_i^2 - \mu_i^1\right)^T \cdot \left(\mu_i^2 - \mu_i^1\right) \cdot \Sigma_i^{-1}\right) \tag{5}$$

where M is the number of components in each model.

As showed in Fig. 5, if the KL distance between X_1 and Y_1 is smaller than that between X_1 and Y_2, X_2 and Y_1, or X_2 and Y_2, speech segments T_1 and T_3 will be merged

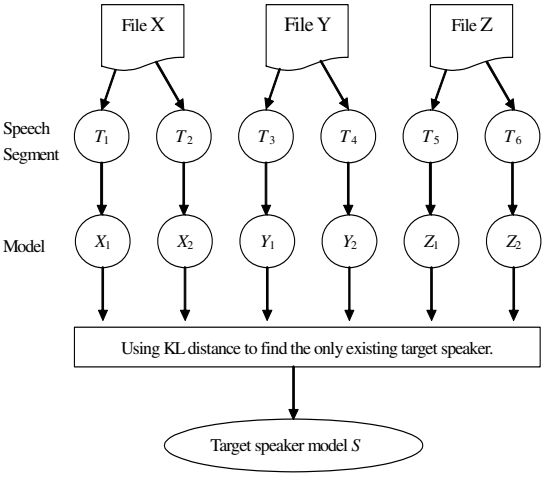

Fig. 5. Multi-speaker training process

together. Finally, a target speaker model S can be obtained from these three 2-spekaer audio files.

The detection results are given in Fig. 6. Comparison result of LIA is showed in Fig. 7 [16].

4.5 Discussion

It can be seen that there exist two large losses: one lies in the performance between 1D and 1T-2D, the other lies in the performance between 2D and 1T-2D. The loss comes from several aspects: (1) there existed many short speech segments and noisy speech segments that might cause errors in segmentation and clustering; (2) there

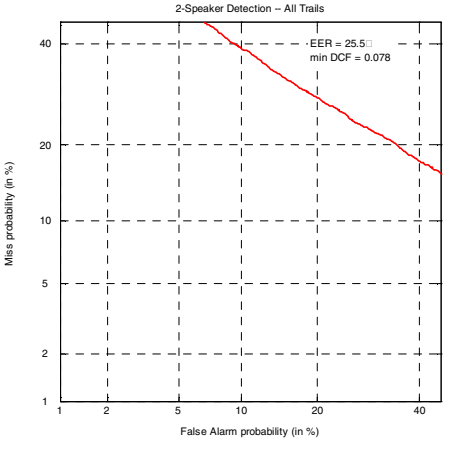

Fig. 6. 2-speaker detection, NIST 2002 evaluation

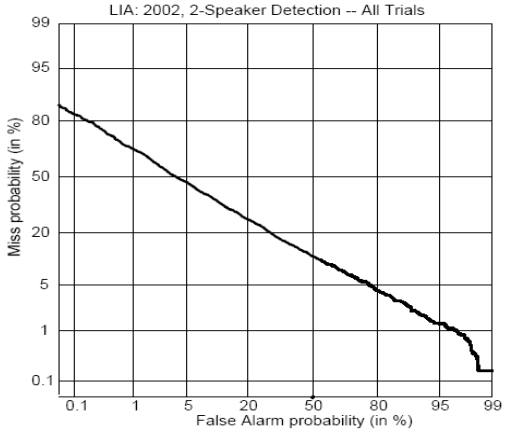

Fig. 7. LIA 2-speaekr result, NIST 2002 evaluation

existed many speech segments spoken by two speakers simultaneously; (3) there existed some mistakes in the multi-speaker training process which might lead to a bad target model; (4) the average duration of speech segments used in 1D was longer than that used in the other two detections; and (5) the errors caused by speaker segmentation can not be corrected by the clustering process.

5 Conclusions

In this paper, a speaker segmentation method based on LLRS over UBM and a speaker clustering method based on difference of log-likelihood scores between two speaker models are proposed. And a complete system with related experiments and results for NIST 2002 two-speaker task is presented. The target models are trained from several multi-speaker speech segments and the tests are also done with 2-speaker files.

The proposed speaker segmentation and clustering methods can achieve a frame error rate of 6.8% on NIST 2002 Switchboard Cellular speaker segmentation corpus. And for 1T-2D, the system achieves an EER of 20.5%, and for 2-speaker detection, the system achieves an EER of 25.5%. The performances of the proposed method on NIST 2002 Switchboard Cellular speaker segmentation corpus, the 1D and 2D tasks are close to that of LIA [16].

Though the segment result seems accurate enough for the task, the performances of 1T-2D and 2D are less satisfactory. Something must be done in order to decrease the detection errors: (1) perform re-segmentation with the speaker models trained in clustering phase; (2) discard the speech segments with bad Signal-to-Noise Ratio (SNR) or overlapped by several speakers; (3) improve the matching strategy during multi-speaker training in order to obtain a more accurate target speaker model.

References

1. Rissanen, J. Stochastic Complexity in Statistical Inquiry. Series in Computer Science, 1989, Vol. 15. World Scientific, Singapore, Chapter 3
2. Chen, S.S., Gopalakrishnan, P.S. Speaker environment and channel change detection and clustering via the Bayesian Information Criterion. In: DARPA Speech Recognition Workshop, 1998
3. Rissanen, J. Stochastic Complexity in Statistical Inquiry. Series in Computer Science, 1989, Vol. 15. World Scientific, Singapore, Chapter 3
4. Gish, H., Siu, M.-H., Rohlicek, R. Segregation of speakers for speech recognition and speaker identification. In: IEEE International Conference on Acoustics Speech and Signal Processing, 1991. 873-876
5. H. Gish and M. Schmidt. Text-independent speaker identification. IEEE Signal Processing Mag. 1994, 11:18-32
6. Siegler, M.A., Jain, U., Raj, B., Stern, R.M. Automatic segmentation classi®cation and clustering of broadcast news audio. In: DARPA Speech Recognition Workshop, 1997. 97-99
7. J. P. Campbell, Jr. Speaker recognition: A tutorial. Proc. IEEE, 1997. 9(85):1437–1462
8. P. Delacourt , CJ Wellekens. DISTBIC: a speaker-based segmentation for audio data indexing, Speech Communication, Sept. 2000, (32):111-126
9. Sylvain Meignier, Jean-François Bonastre, and Stéphane Igounet. E-HMM approach for learning and adapting sound models for speaker indexing. In 2001: A Speaker Odyssey, Chania, Crete, June 2001. 175-180
10. D. Moraru, S. Meignier, L. Besacier, J.-F. Bonastre, Y. Magrin-Chagnolleau. The ELISA consortium approaches in speaker segmentation during the NIST2002 speaker recognition evaluation, In Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2003), Hong Kong, 2003. (2): 89–92
11. D. Moraru, S. Meignier, C. Fredouille, L. Besacier, J.-F. Bonastre. The ELISA consortium approaches in broadcast news speaker segmentation during the NIST2003 rich transcription evaluation, In Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2004), Montreal, Canada, 2004
12. T. Wu, L. Lu, K. Chen, and H. Zhang. UBM-based real-time speaker segmentation for broadcasting news. In Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing（ICASSP2003）, Hong Kong, China, 2003. (2):193–196
13. D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing. 2000, (10):19-41
14. Zhenyu Xiong, Thomas Fang Zheng, Zhanjiang Song, and Wenhu Wu. Combining Selection Tree with Observation Reordering Pruning for Efficient Speaker Identification Using GMM-UBM. Proc. ICASSP. 2005, 625-628
15. http://www.nist.gov/speech/tests/spk/2002/resource/index.htm
16. Jean-François Bonastre, Sylvain Meignier, Teva Merlin. Speaker detection using multi-speaker audio files for both enrollment and test. In ICASSP 2003, Hong Kong, China