A NEW SPEAKER CHANGE DETECTION METHOD FOR TWO-SPEAKER SEGMENTATION

André G. Adami[1], Sachin S. Kajarekar[1], Hynek Hermansky[1,2]

[1]OGI School of Science and Engineering, Oregon Health and Science University, Portland, USA
[2]International Computer Science Institute, Berkeley, California, USA
{adami, sachin, hynek}@asp.ogi.edu

ABSTRACT

In absence of prior information about speakers, an important step in speaker segmentation is to obtain initial estimates for training speaker models. In this paper, we present a new method for obtaining these estimates. The method assumes that a conversation must be initiated by one of the speakers. Thus one speaker model is estimated from the small segment at the beginning of the conversation and the segment that has the largest distance from the initial segment is used to train second speaker model. We describe a system based on this method and evaluate it on two different tasks: a controlled task with variations in the duration of the initial speaker segment and amount of overlapped speech and 2001 NIST Speaker Recognition Evaluation task that contains natural conversations. This system shows significant improvements over the conventional system in absence of overlapped speech on the controlled task.

1. INTRODUCTION

The goal of speaker segmentation is to obtain speech segments spoken by each speaker in a conversation. These segments can be used for speaker adaptation in speech and speaker recognition systems. This is a difficult task in absence of prior information about speakers. It becomes even more challenging when the number of speakers in a conversation is not known, and the speakers speak simultaneously. In this paper, the conversations have two speakers that may speak simultaneously.

The two-speaker segmentation can be divided into four steps: feature extraction, speaker change detection, clustering, and resegmentation. The feature extraction converts the speech conversation into some parameterized representation. The speaker change detection step splits the conversation into smaller segments that are assumed to contain only one speaker. The clustering step merges all the segments until two clusters remain and the speaker models are estimated from each cluster. Finally, the resegmentation step to performs a more refined segmentation using these speaker models.

The speaker detection step is the most important part in the segmentation because the segments produced from this step are used to estimate the speaker models. This means that if the speaker detection step produces segments that contain more than one speaker then the speaker models will be estimated incorrectly. Therefore, we are investigating into a new speaker change detection method.

Two approaches commonly used for speaker change detection are energy-based [1][2] and distance-based [3][4][5] and. The second approach [1][2] assumes that the probability of a speaker change is higher around silence regions. It uses speech-silence detector to identify the speaker change locations. The distance-based method searches for the speaker change candidates at the maxima of the distances computed between adjacent windows over the entire conversation. The hypothesized speaker changes are validated based on a threshold that is shown to vary across different conditions [3].

This paper presents a new speaker change detection method for two-speaker segmentation that requires neither a threshold nor existence of silence regions in the conversation. It assumes that one speaker initiates the conversation and he/she speaks for at least one second. Section 2 describes this method for two-speaker segmentation. Section 3 describes the National Institute of Standards and Technology (NIST) evaluation setup and the database of conversations created from HTIMIT [6]. Section 4 presents the description of the speaker segmentation systems. Section 5 presents the results of the speaker segmentation systems on the NIST database and the database created from HTIMIT.

2. PROPOSED SPEAKER CHANGE DETECTION METHOD

When we listen to a conversation, we know when a speaker change occurs even whether there is no silence between different speakers or the speech is overlapped. Thus, we can assume that at the first moment that we

listen to someone's voice, we memorize it to use for future references in the conversation. It is also true to say that any type of conversation must be initiated by one of the participants. Using these assumptions, we have developed a speaker change detection algorithm that uses some data from the conversation to find the regions where each speaker is speaking. The algorithm is described as follows:

1. **Speaker 1 data selection**: the beginning of the conversation (segment S_1 in Figure 1(a)) is assumed to represent speaker 1. The size of the segment is one second, because we assume that he/she speaks for at least one second;
2. **Distance Computation**: a sequence of Generalized Likelihood Ratio (GLR) distances [5] is computed between the data selected for speaker 1 (S_1) and shifted segments (d_i in Figure 1(a)) over the conversation. The distance sequence is mean and variance normalized and smoothed using a sliding window;
3. **Speaker 2 data selection**: the segment d_i, whose distance is the largest and is not silence, is assumed to represent speaker 2. The arrows in Figure 1(b) point out the candidates regions to represent speaker 2;
4. **Distance Computation**: using the selected region for speaker 2, a new sequence of distances is computed between that region and shifted segments over the conversation;
5. **Segment Assignment**: the segment boundaries are defined in the points where the distances with respect to both speakers are equal (as shown in Figure 1(c)). Each segment in the conversation is assigned the speaker with the smallest distance;
6. **Segmentation Refinement**: after segment assignment, we have a better estimate of the speaker segments. Therefore, more segments can be selected for each speaker where the difference between the distances from both speakers is the highest (the double arrows in Figure 1(c) point out some candidates segments). Then, the steps from 1 to 5 can be repeated.

3. EVALUATION SETUP

In this section, we describe two databases for evaluating the speaker segmentation systems and the scoring process. The development database has conversations created by concatenating speech samples from HTIMIT and the NIST Speaker Segmentation database is composed of natural conversations. The former allows us to study the effects of variation in the overlapped speech and the speech duration in the beginning of the conversation.

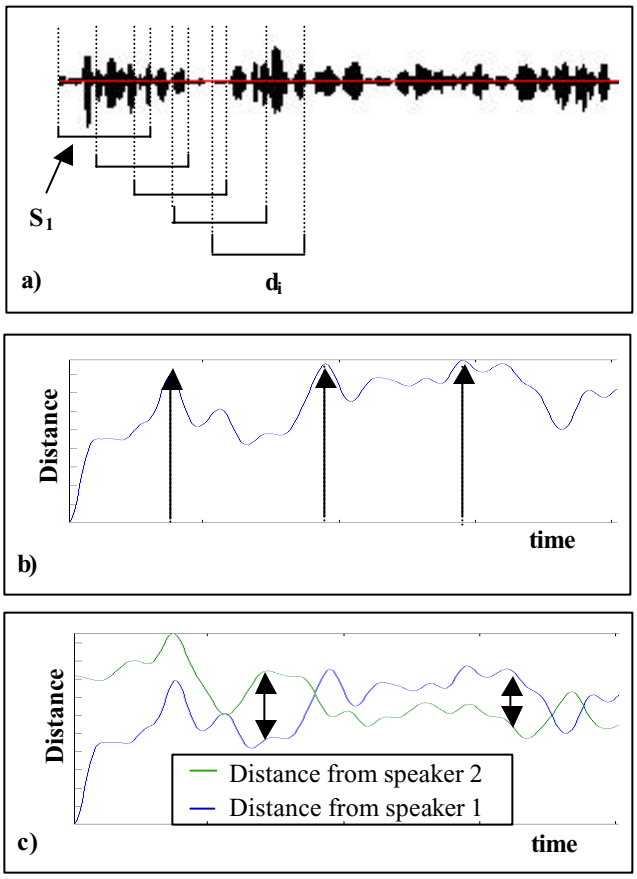

Figure 1 – Steps of the speaker change detector: a) distance computation, b) second speaker data selection, and c) new data selection for both speakers.

3.1. Development Database

We have used HTIMIT [6] database to create conversations for the segmentation task. HTIMIT is a re-recording of part of TIMIT corpus through 10 different telephone handsets. Each conversation in the development database is a concatenation of speech segments from two speakers over 4 different electret handsets. This results in 720 one-minute conversations with 20 speaker turns on average.

The database is divided into 4 conditions:

1. No overlapped speech and single speaker speaking at least one second in the beginning of the conversation (C1);
2. No overlapped speech and single speaker speaking less than one second in the beginning of the conversation (C2);
3. Overlapped speech and single speaker speaking at least one second in the beginning of the conversation (C3), and
4. Overlapped speech and single speaker speaking less than one second in the beginning of the conversation (C4);

Each condition has 60 conversations between male speakers, 60 between female speakers, and 60 between genders.

3.2. NIST Speaker Segmentation Database

The database used in the NIST Evaluation [7] consists of 1000 telephone conversations drawn from the Switchboard-2 Corpus Phase II. Each conversation is about 60 seconds long and has regions of simultaneous speech. There are 269 conversations among male speakers, 323 conversations among female speakers, and 408 conversations between male and female speakers.

3.3. Scoring Process

The output of the speaker segmentation systems must be the time intervals during which each speaker is speaking in a conversation. The hypothesized time intervals are compared to reference time intervals using the NIST scoring tool [7]. The final score is the classification error, which is obtained by the ratio between the amount of corrected label speech (CLD) and the total amount of speech (TS): 1 – CLD/TS. Since overlapped speech regions belong to both speakers, they are not taken into account in the scoring process.

4. SYSTEM DESCRIPTION

Most of the speaker segmentation systems use Mel-frequency Cepstral coefficients [1][2][3][4][5]. However, in previous experiments using the development database, Line Spectral Pair has shown around 20% improvement over the Mel-frequency Cepstral coefficients. Therefore we use 24 LSP [8] coefficients as features for both systems. They are computed every 10ms using a 32ms Hamming window.

4.1. System using Proposed Method

For speaker change detection, a one-second segment is selected as a reference for both speakers in the first pass. A 4 component Gaussian Mixture Model (GMM) with diagonal covariance is used to compute the GLR distance. Each distance is computed over one second window and the window is shifted by 0.1 second. The sequence of distances is smoothed using a 2-second hamming window. In the second iteration the size of the segment for each speaker is increased to 3 seconds and a 8 component GMM with diagonal covariance is used to compute the GLR distance. This system performs only two iterations because the performance does not increase significantly beyond the second iteration.

The segments created by the speaker change detection step are used to initialize the clustering algorithm. This algorithm is an agglomerative method that computes the GLR distance between every pair of clusters and merges two clusters with the minimum distance at every step. The GLR distance is computed using a 16 component GMM with diagonal covariance. Clustering is repeated until two clusters are formed.

The resegmentation step is performed as follows. First, a 32 component GMM with diagonal covariance (background model) is trained using the entire conversation. Using the data from each cluster, the speaker specific models are adapted from this GMM using Maximum A-Posteriori training [1]. Then, the likelihood ratio score between the speaker models and the background model is computed for each frame. The score sequences are smoothed using a 2.5 second hamming window. Finally, the frames are assigned to the speaker with the highest likelihood score.

4.2. System using an Energy-based Method

The system is based on the energy-based speaker change detection method described in [2]. This method hypothesizes that a speaker change is more likely to occur around silence regions. It uses an adaptive energy-based speech-silence detector to create one-second speech segments. The clustering and resegmentation steps are the same as described in section 4.1.

5. RESULTS

Table 1 shows the results for the systems on the development database.

Table 1 –Systems Error Rate on the Development Database

Condition			Energy-based Method	Proposed Method
	Overlapped Speech	Initiator Segment Duration		
C1	N	1.0 sec	0.06	0.02
C2	N	0.5 sec	0.07	0.04
C3	Y	1.0 sec	0.08	0.05
C4	Y	0.5 sec	0.09	0.09

Table 1 shows that the system using the proposed method performs better than system using the energy-based method for the conversations without overlapped speech (C1). The proposed method gives similar performance without the clustering step, which suggest that the clustering step is more important for the energy-based method than the proposed method.

The performance of the proposed method decreases when the conversation initiator speaks for less than one second. This is expected because it is assumed that the conversation initiator speaks for at least one second.

However, the proposed method still performs better than the energy-based method.

None of these methods employ special processing for the overlapped speech. Therefore, the conversations with overlapped speech (C3) affect both the proposed and the energy-based methods. Notice that the performance deterioration in the proposed method is more than the one caused by conversations where the initiator speaks less than one second (C2). The effect of overlapped speech and the initiator speaking for less than one-second (C4) is not additive. The result shows that both systems perform the same under this condition.

Table 2 presents the performance of the segmentation systems on the NIST database. The results show that both methods perform comparable on the NIST task. Note that this is similar to the C4 condition from the development database. However, the better performance on this database can be attributed to the fact that 66% of the conversations have the initiator speaking for at least for one second.

Table 2 – Systems Error Rate on the NIST Database

Method	Error
Energy-based Method	0.07
Proposed Method	0.07

6. CONCLUSIONS

In this paper, we proposed a new speaker change detection method for two-speaker segmentation. This method assumes that one speaker will initiate the conversation and the speaker speaks for at least one second. A speaker segmentation system using the proposed method was evaluated on two databases: 1) controlled database created from HTIMIT and 2) NIST Speaker Segmentation database.

The system based on the proposed method performed better than the baseline on conversations without overlapped speech. Under the overlapped speech condition, the proposed method performs better if the conversation initiator speaks for at least one second. The proposed method performs the same as the energy-based method on the conversation with overlapped speech and initiator speaking less than one-second.

For future work, we plan to study different feature sets for the proposed system. Another issue is to modify the method to overcome problems like noise in the beginning of the conversations and the overlapped speech. We also plan to extend this method to N-speaker segmentation task, which does not make any assumption about the number of speakers.

7. ACKNOWLEDGEMENTS

Authors would like to thank Kirk Jackson for useful discussions on this work during his visit to OGI. This research was funded by DoD under grant MDA904-00-C-2089.

8. REFERENCES

[1] T. Kemp, M. Schmidt, M. Westphal, A. Waibel, "Strategies for Automatic Segmentation of Audio Data," Proc. of ICASSP-2000, pp. 1423-1426, 2000.

[2] D. A. Reynolds, R. B. Dunn, J. J. McLaughlin, "The Lincoln Speaker Recognition System: NIST EVAL2000", Proc. of ICSLP-2000, Beijing, 2000.

[3] P. Delacourt, C. J. Welkens, "DISTBIC: A Speaker-based Segmentation for Audio Data Indexing," Speech Communication, v. 32, pp 111-126, 2000.

[4] S. Chen, P. Gopalakrishnan "Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion", DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne, VA, 1998.

[5] H. Gish, M. Siu, R. Rohlicek, "Segregation of Speakers for Speech Recognition and Speaker Identification," Proc. of ICASSP-91, pp. 873-876, 1991.

[6] D. Reynolds, "HTIMIT and LLHDB: Speech Corpora for the Study of Handset Transducer Effects," Proc. of ICASSP-97, pp. 1535-1538, 1997.

[7] National Institute of Standards and Technology, "The 2001 NIST Speaker Recognition Evaluation," http://www.nist.gov/speech/tests/spk/2001, 2001.

[8] F. Itakura, "Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals," J. Acoust. Soc. Am., vol 57, S35, 1975.