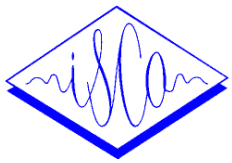

7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

SPEAKER CHANGE DETECTION USING A NEW WEIGHTED DISTANCE MEASURE

Soonil Kwon, Shrikanth Narayanan

Department of Electrical Engineering, Speech Analysis and Interpretation Lab, and Integrated Media Systems Center
University of Southern California
soonilkw@usc.edu, shri@sipi.usc.edu

ABSTRACT

Speaker change detection is a key pre-requisite to speaker tracking and speaker adaptation. It detects the points where a speaker identity changes in a multi-speaker audio stream. We first extract the speech segments from an audio stream by segmentation and classification techniques. Using the extracted speech segments, the proposed weighted metric-based technique detects the speaker change points. New weights are originated from Fisher Linear Discriminant Analysis and, when used with Mel Cepstrum feature vectors, it has an effect of subband processing. Experiments were performed with HUB-4 Broadcast News Evaluation English Test Material (1999) and a movie audio track. Results showed that our technique gave about 37.7% improvement compared with Euclidean distance on the broadcast news data and about 27.1% on the movie data; with Mahalanobis distance, the improvements were 37.7% and 25.3% for broadcast news and movie data, respectively.

1. INTRODUCTION

Automatic segmentation and classification of an audio stream according to speaker identities and environmental conditions are gaining increasing importance as we get more and more information from TV programs and movies [6]. For example, this process is useful in the task of automatic transcription and indexing of broadcast news or movie audio data, speaker adaptation techniques for advanced speech recognition systems, and speaker tracking in multimedia data processing. For unsupervised transcription or adaptation systems, more data from the same speaker does help to reduce error rates significantly [4]. In this paper, we propose a method for unsupervised speaker change detection and evaluate it with broadcast news data and movie audio data.

The goal of unsupervised speaker change detection is to detect the points where the speaker changes in the middle of an audio stream without any knowledge about the identity and number of speakers. The speaker change detection system consists of three steps. The first step is segmentation; the audio data contains clean speech, speech with background music or noise, music, noise, etc. An audio stream is divided into segments that are 'homogeneous' regions with respect to audio characteristics. For example, a segment contains only speech, music, or speech and music. The second step is classification. We next classify segments, which are obtained from the first step, into several broad categories. Our classification is based on the combination of a silence ratio and a zero-crossing variation [3]. From the results of classification, we collect all segments classified as containing speech including those with other music and environmental noise. In the third step, the system sequentially detects whether the speaker changes between two neighboring speech segments.

The primary focus of this paper is on the third step. The first two steps, although not completely solved problems, have been well studied in the literature. We adopt and expand on methods that have resulted in state of the art performance for these steps. There are three general speaker detection techniques: metric-based, model-based, and decoder-guided [6]. A metric-based algorithm calculates the distance between neighboring segments. Even if it relies on the threshold of measurements, it is simply applied without a large training data set and prior knowledge of speakers. Nonprobabilistic distances, such as Euclidean distance and Mahalanobis distance, measure the dissimilarity between two feature vectors. The distance decreases as the similarity between two feature vectors increases. For a given distance, we can also define the corresponding weighted version by introducing a weighting factor in the distance equation [1].

The main idea of this paper is a new weighted distance measure that augments previous distance measure techniques. We propose this new weighted measure to improve the performance of discriminating speakers. The usual weight with the variance of feature within a class gives smaller weights to some features with large variances and larger weights to some features with small variances. This technique is augmented by interclass information, which also has an important role in classification. When we use Mel Cepstrum as feature vectors, this weighted distance measure has an effect of subband processing and makes it possible to implement an unsupervised detection.

We experimented with a part of HUB-4 Broadcast News Evaluation English Test Material (1999) and a movie audio track from 'When Harry met Sally' (1989). We obtained the thresholds from about 5 minutes of data for broadcast news and 10 minutes of data for movie audio. The experimental results show that, when compared with Euclidean distance, the precision of speaker change detection was improved up to 37.7% on the broadcast news data and 27.1% on the movie audio track; with Mahalanobis distance, the improvements were 37.7% and 25.3% for broadcast news and movie audio, respectively. Improvements provided by our speaker detection algorithm, in turn, can enhance unsupervised speaker indexing or adaptation systems.

In section 2 of this paper, details of the unsupervised speaker change detection algorithm will be described. We will

then explain how our new weight is calculated and what aspect of this weight improves the performance of unsupervised detection in section 3. In section 4, we will give an explanation of our experiment, and, in section 5, the experimental results will be shown and discussed. In section 6, the conclusion of this paper will be presented.

2. SPEAKER CHANGE DETECTION

2.1. Segmentation of audio data

The general goal of segmentation is to get a sequence of discrete utterances from an audio stream. Each segment has a certain characteristic: music, noise, or speech. Sometimes a segment may have two or more such characteristics.

There are two kinds of segmentation: fixed length segmentation and variable size segmentation. In fixed length segmentation, audio data is chopped into segments, the length of which is predetermined. Fixed length segmentation is simple, but relatively long segments are likely to include speaker changing points, while short segments do not have enough speaker information. In variable length segmentation, the audio data is divided into variable length of segments.

In the middle of speaking, people usually breathe. There is low probability of a speaker changing between breathing points. In our unsupervised variable size segmentation, audio data is segmented by silence points in the middle of the sequence of speech. The silence point is defined as a certain period within which the energy of a signal stays below the threshold. To avoid potential problems of false alarm, some restrictions are placed on the length of segments. If the length of a segment is shorter than a threshold, it is merged with neighboring segments depending on the length of silences [2]. For example, if front-end silence of a segment is longer than rear-end silence, then the segment is merged with the next segment. The period and threshold can be determined depending on the characteristics of the data sources.

2.2. Classification of segments

Generally audio data can be categorized into five broad classes: music, noise, clean speech, speech with music, and speech with noise. While we need only speech segments, it is very difficult to separate speech from background music or noise. For this reason, the speech class includes the speech with music class and the speech with noise class. In this stage, we just classify speech signals and non-speech signals with a silence ratio and the level of variation in the zero crossing rates [3].

Speech signals have a higher silence ratio than music. We determined two thresholds for calculating the silence ratio. One is for the amplitude, which indicates small energy. The other is for the length of period of low energy over which the signal is considered as silence. We find the total length of silence and divide it by the length of a segment for normalization. Every threshold is empirically obtained from experiments.

It is well-known that speech has a higher level of variation in zero crossing rates. The variation of zero crossing rate of each segment is calculated and normalized.

The primary purpose of this stage is extracting speech signals. However, noise signals include various types, such as babbling, that may contribute to increased errors. Since it is critical that we should not lose any speech segments, the focus of the classification is to minimize false rejection even at the cost of false acceptance.

2.3. Speaker change detection

When the identity and number of speakers are unknown, we need an unsupervised detection technique. Since our unsupervised speaker change detection technique is metric based, it does not need any knowledge or models of speakers and a large training data set.

After the data segmentation and classification procedure, speech segments are ready for speaker change detection. Within each segment, there is no speaker change because we assume that the probability of the speaker changing between breathing points is small. Therefore we can perform speaker detection at the segment level.

Speech signals are converted into feature vectors and then the distance between two neighboring segments is sequentially calculated. If the distance is above a threshold, then the point between two neighboring segments is assigned to a speaker changing point. Thresholds are determined by using a small number of the data segments (as explained in the section 4).

3. NEW DISTANCE MEASURE

The simple Euclidean distance classifier is one of the fundamental and widely used techniques in classification. We use Euclidean distance as the basis of our new weighted distance.

3.1. Weighted squared Euclidean distance

Let $X^{(k)} = \{x_1^{(k)}, x_2^{(k)}, ..., x_{n_k}^{(k)}\}$ be the sample space in class k (k = 1, 2, …,C) where n_k is the number of samples in class k. $x_i^{(k)} = (x_{i1}^{(k)}, x_{i2}^{(k)}, ..., x_{im}^{(k)})$ is an i-th feature vector (i = 1, 2,…, n_k) where $x_{ij}^{(k)}$ is the j-th feature in the i-th sample (j = 1, 2,…, m). In our case, since two neighboring segments are compared at a time, the number of classes, C, is 2. The mean vector of class k is

$$\bar{x}_j^{(k)} = (\sum_{i=1}^{n_k} x_{ij}^{(k)}) / n_k \quad . \tag{1}$$

The Euclidean distance between two neighboring segments is defined as

$$d_e = \sqrt{(\sum_{j=1}^{m} (\bar{x}_j^{(1)} - \bar{x}_j^{(2)})^2) / m} \quad . \tag{2}$$

The intraclass variances of features within each class or the interclass variance between classes affects the performance of a classifier. Weighting techniques show better results for increasing the similarity of the samples in the same class and decreasing the similarity of the samples from different classes [1].

We propose a weighted squared Euclidean distance (WED)

$$d_{ki} = \sum_{j=1}^{m} w_j^{(k)} (\bar{x}_j^{k} - \bar{x}_j^{i})^2 \tag{3}$$

with two constraints:

$$w_j^{(k)} \geq 0 \quad , \quad \sum_{j=1}^{m} w_j^{(k)} = 1 \quad . \tag{4}$$

3.2. New weights

The usual weight is the variance of feature within a class. It gives smaller weights to some features with large variances and larger weights to some features with small variances. This technique makes the importance of each feature equal within a class. However, it reflects only intraclass information of features. Interclass information also plays an important role in classifiers. For that reason, we use the variance of two classes, which are from two neighboring segments in our case, as weights [1]. This idea is originally from Fisher Linear Discriminant Analysis. Since between-class scattering information has computational difficulty, it is replaced by total scattering information, which is the sum of a within-class scatter value and a between-class scatter value.

We define new weights as

$$w_j' = \left(\frac{\text{t var}_j}{\text{w var}_j}\right) \Big/ \left(\sum_{j=1}^{m} \frac{\text{t var}_j}{\text{w var}_j}\right) \quad (5)$$

where tvar_j is the variance of total j-th feature vectors from two neighboring segments and wvar_j is the sum of the variance of j-th feature vectors from segment 1 and the variance of j-th feature vectors from segment 2.

Mel Cepstrum feature vectors are used for our unsupervised speaker change detection technique. $x_{ij}^{(k)}$, which is the j-th feature in i-th sample, is from the different frequency bands of the filter bank. Each weight is calculated with features from the same frequency band, which implies that each feature from a different frequency band is weighted differently depending on the within-class variance and between-class variance. Therefore it provides the same effect as subband processing.

For improving the performance of discrimination, we apply a Sigmoidal function to new weights [Eq.(5)]. The Sigmoidal function provides a larger discriminative power near the threshold of decision. It spreads out the range of weights with respect to the mean of weights. The parameter of Sigmoidal function [Eq.(6)], p, is always positive and, as it is larger, the curve of Sigmoidal function is steeper, which makes the spreading range larger (*Figure 1*).

$$w_j = \frac{1}{1 + e^{-p(w_j' - \overline{w'})}} \quad (6)$$

where $$\overline{w'} = \sum_{j=1}^{m} w_j' \Big/ m \ . \quad (7)$$

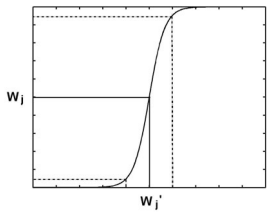

Figure 1: Sigmoidal function

4. EXPERIMENTS

In this experiment, we used two kinds of audio data: broadcast news and audio data from a movie. Sixty minutes of the broadcast news data was extracted from HUB-4 Broadcast News Evaluation English Test Material (1999) and 20 minute movie audio data was from 'When Harry met Sally' (1989). Both data were sampled at 16 kHz (16 bit and mono).

There were 3 steps in our experiment. The first step was a segmentation process. An unsupervised segmentation process divided the broadcast news data into 669 segments and the movie data into 480 segments. All types of audio data were included, but segmentation was performed based on speech data. Therefore, there were 668 segment transitions in the broadcast news data and 479 segment transitions in the movie data. In the segmentation process, the threshold for the length of segments was determined experimentally. In the broadcast news data, we preset the time threshold as 2 seconds to avoid potential problems of false detection due to segments too short to include enough information of speakers while, in the movie data, it was 0.5 seconds. In movies, the frequency of speaker change in a conversation is higher than in news. In addition, in the movie data, it was more likely for a speaker to interrupt another speaker.

The second step was an audio classification process. Both audio data were classified into 3 types: speech, music, and noise. The speech class included clean speech, speech with music, and speech with background noise. The music class included pure music without speech, but it could include singing voices. The noise class included only noise without speech.

The third step was an unsupervised speaker change detection process. The distance threshold was preset by the lowest error rate that was obtained from 5 minutes of data for the broadcast news and 10 minutes of data for the movie audio data regardless of the length of full data for testing. In our experiments, to implement this, a 15 minute chunk of data was chosen arbitrarily (from each domain) and was systematically reduced with the estimation of error variance each time. The smallest amount of data that yielded about 1% error variance was deemed 'optimal' for that domain. For the broadcast news domain, it was about 5 minutes while, for the movie audio, it was 10 minutes. The reason for the larger amount of data for the movie domain is because it had a more nonstationary environment than the broadcast news domain.

Feature vectors, 48 channel and 32 dimensional Mel Cepstrum, were obtained from speech segments for speaker change detection. We experimentally obtained the parameter of Sigmoidal function: 150 for the broadcast news data and 200 for the movie data.

5. RESULTS AND DISCUSSION

We segmented audio streams with an assumption that there is seldom a speaker change within a segment that is located between breathing points. Unsupervised segmentation introduced two kinds of errors: one arose from multiple speakers in a segment and the other from background music or noise. The first type of error resulted from the speech segment being too short for segmentation or due to interruption of another speaker in the middle of speaking. The second type of error came from significant changes in background noise or music in the middle of a speech from one speaker. It also resulted when some noise or music was present in the middle of a silence region separating two different speakers.

Segments were categorized into three classes: music, noise, and speech. The speech class included all speech segments with any background condition. Due to the characteristics of movies,

most of the speech segments include background music, singing, or noise that caused speaker change detection errors. In the broadcast news data, most of the speech segments were not corrupted by background noise or music. Even though we included the characteristics of these two kinds of data in the modeling experiments, we still got significantly better performance on the relatively clean broadcast news data (*Table 1*).

		Broadcast News	Movie
Total Segment		669	480
Noise or Music Segment		14	58
Speech Segment		573	330
Segmentation Error Counts	Type 1	64	81
	Type 2	18	11
	Total	82	92
Segmentation Error Rate		12.3 %	19.2 %

Table 1: Segmentation and Classification

Metric	Broadcast News	Movie
Euclidean	8.4 (37.7%)	21.6 (27.1%)
Mahalanobis	8.4 (37.7%)	21.3 (25.3%)
Squared Euclidean	6.8 (11.5%)	20.0 (17.6%)
Weighted Squared Euclidean	6.1	17.0

Table 2: Error Rate (%) of Speaker Change Detection. Relative improvement is shown parenthetically with respect to the new metric.

Table 2 summarizes the results for the unsupervised speaker change detection. The error in *Table 2* is the sum of false acceptance errors and false rejection errors. Since the weight was only based on the within-class scatter, Mahalanobis distance measure could not reflect the dissimilarity between classes. But weighted squared Euclidean distance measure gave better results due to the new weights, which reflected within-class and between-class information.

Our new metric provided about 37.7% improvement compared with Euclidean distance and Mahalanobis distance for the broadcast news data and about 27.1% with Euclidean distance and 25.3% with Mahalanobis distance for the movie data (*Table 2*). We preset the time threshold as 2 seconds for the broadcast news data but, for the movie data, it was 0.5 seconds. Segmentation resulted in many short speech segments (shorter than 1 second) for the movie data. Many errors of speaker change detection were due to these short speech segments in our experiment with movie audio data. However, a larger time threshold yielded the larger number of segments including multiple speakers, which increased segmentation errors.

6. CONCLUSION

We presented a new weighted metric-based technique for unsupervised speaker change detection with broadcast news data and movie audio data. With an assumption that there is no speaker change between breathing points, the audio data was segmented by silence points in the middle of the speech sequence. The silence point was defined as a certain period within which energy of a signal stays below the threshold. This segmentation technique lowered the number of segments processed without losing the merit of a short fixed segment method. However, it still has problems in noisy environments. To solve this problem, an adaptive segmentation algorithm should be adopted.

Audio classification of our experiment correctly classified most of the segments. One of the most difficult problems was to classify the speech with background music from a background singing voice with music. More features may give us better performance.

Our weights included not only intraclass information of features but also interclass information. Since each feature from a different frequency band was differently weighted depending on the within-class variance and between-class variance, weighted distance measurement made a more accurate detection. In experiments, even though some speech segments were corrupted by various high energy level background noise and music, the precision of detection was improved up to 37.7% (*Table 2*). This result means that our weighted distance measure discriminated the speakers more precisely without any knowledge of speakers and a large training data set. However, there were two problems for the speaker change detection. The first problem was that a speaker segment with two or more kinds of background noise or music could not be correctly detected. It was not determined whether it could be considered as one speaker segment or not. The second problem was that there were still short segments that do not contain enough information for detection. There is no obvious solution for this problem at this time.

To further improve the overall performance of the unsupervised speaker change detection, we need a more robust segmentation algorithm followed by an adaptive distance measure technique.

7. REFERENCES

[1] Lin, H. and Venetsanopoulos, A.N., "A Weighted Minimum Distance Classifier for Pattern Recognition", Canadian Conference on Electrical and Computer Engineering, vol.2, 904-907, 1993.

[2] Hain, T., Johnson, S.E., Tuerk, A., Woodland, P.C., and Young, S.J., "Segment Generation and Clustering in the HTK Broadcast News Transcription System", Proc. DARPA Broadcast News Transcription and Understanding Workshop, 133-137, 1998.

[3] Lu, G. and Hankinson, T., "An Investigation of Automatic Audio Classification and Segmentation", Proceedings of ICSLP2000, 776-781, 2000.

[4] Magrin-Chagnolleau, I. and Bimbot, F., "Indexing Telephone Conversations by Speakers Using Time-Frequency Principal Component Analysis", IEEE Internatinal Conference on Multimedia and Expo 2000, Volume: 2, 881-994, 2000.

[5] Nishida, M. and Ariki, Y., "Speaker Indexing for News Articles, Debates and Drama in Broadcasted TV Programs", IEEE International Conference on Multimedia Computing and Systems 1999, Volume: 2, 466-471, 1999.

[6] Chen, S.S. and Gopalakrishnan, P.S., "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion", DARPA Broadcast News Transcription and Understanding Workshop, 127-132, 1998.