Speaker-and-environment Change Detection in Broadcast News using the Common Component GMM-based Divergence Measure

Yih-Ru Wang and Chi-Han Huang
Department of Communications Engineering, National Chiao Tung University, Taiwan, Republic of China
yrwang@cm.nctu.edu.tw

Abstract
In this paper, a GMM with common mixture components, referred to as the common component GMM (CCGMM), is proposed to be the signal model for calculating the diversity measure for the speaker-and-environment change detection in broadcast news signal. The use of GMM is to increase the accuracy of audio signal modeling while the use of common mixture components is to solve the complexity problem of parameter estimation and similarity measure evaluation. Experimental results on a TV broadcast news database showed that it outperformed a BIC-based method. A MDR of 21.9% with 16.0% FAR was achieved.

1. Introduction
The segmentation of broadcast news signal is an important technology because large amounts of information were delivered through the broadcast news everyday. A good segmentation scheme is useful for further processing to categorize, archive and retrieve the information in broadcast news. Usually, an unsupervised speaker-and-environment change detection will be the first phase, and then the speaker/story segmenting and tracking can be done based on the candidates found from the speaker-and-environment change detection. The speaker-and-environment change detection is a difficult task for audio broadcast news signals because they contain many diverse sources including speech of different speakers with different speaking styles, background noises and non-speech signals. The recording conditions of speakers like anchors inside studio, reporters and interviewees in the field, and speech in advertisement are totally different. And, lots of background signals, such as music, noise and chat, may appear in the broadcast news.

A general way to solve the problem of speaker-and-environment change detection in broadcast news is to find a good measure which can indicate the statistics similarity/dissimilarity between two audio segments beside a candidate point. Many similarity measures of audio signal were proposed in the past. They included Kullback Leibler (KL) distance, symmetric Kullback Leibler distance (KL2) [1], divergence shape distance [2,3], and Bayesian information criterion (BIC) [4,5]. They usually assumed that the probability density function of some features of the audio signal in a 2- to 3-second interval was Gaussian. Although it seems that the assumption is too rough for a 2- to 3-second audio segment, a more precise signal model, like the Gaussian mixture model (GMM), may cause the complexity issue in parameter estimation and similarity measure evaluation.

In this paper, a GMM with common mixture components, referred to as the common component GMM (CCGMM), is proposed to be the signal model for calculating the diversity measure for the speaker-and-environment change detection in broadcast news signal. The use of GMM is to increase the accuracy of audio signal modeling while the use of common mixture components is to solve the complexity problem of parameter estimation and similarity measure evaluation.

The paper is organized as follows. Section 2 describes the proposed CCGMM-based divergence measure for speaker-and-environment change detection. Section 3 discusses the experimental results for a television broadcast news database. Some conclusions are given in the last section.

2. The CCGMM-based divergence measure
The divergence measure [6] is usually used to measure the dissimilarity of two random variables based upon the information theory. It is defined as the average of the discrimination of two classes and can be expressed by

$$D = \frac{1}{2} \int \left( \ln \frac{p_1(x)}{p_2(x)} \right) \frac{1}{2} \ln \left( \frac{p_1(x)}{p_2(x)} \right) dx$$

where $p_1(x)$ and $p_2(x)$ are the probability density functions of the two random variables which can be two audio signals in the speaker-and-environment change detection problem.

The distributions of both random variables are assumed to be Gaussian, i.e.,

$$p_1(x) = N(x | \mu_1, \Sigma_1), \quad p_2(x) = N(x | \mu_2, \Sigma_2),$$

where $\mu_i$ and $\Sigma_i$ are the mean vector and the covariance matrix of the $i$th distribution. Then, the divergence can be simplified as

$$D = \frac{1}{2} \left( \Sigma_1 - \Sigma_2 \right) \left( \Sigma_1^{-1} - \Sigma_2^{-1} \right) \left( \mu_1 - \mu_2 \right)^T \left( \Sigma_1^{-1} - \Sigma_2^{-1} \right) \left( \mu_1 - \mu_2 \right)$$

Usually only the first term in Eq. (3), which is called the divergence shape distance, is used to measure the similarity of two audio segments in broadcast news segmentation [2]. But, the Gaussian distribution assumption is very rough for the feature vectors collected from 2-3 seconds audio signal. If a mixture Gaussian distribution is used to substitute the Gaussian distribution, i.e.,
The above equation can be simplified by using common component GMM (CCGMM) models. In a CCGMM, a set of Gaussian distribution components, \( N(\mathbf{x} | \mathbf{\mu}_n, \mathbf{\Sigma}_n) \) \( \forall n=1, \ldots, N \), are used. Thus

\[
p_i(x) = \sum_{n=0}^{N-1} c_{in} N(\mathbf{x} | \mathbf{\mu}_n, \mathbf{\Sigma}_n) \quad \forall i=1,2, \ldots, (4)
\]

then the divergence measure becomes

\[
D = \left[ \sum_{n=0}^{N-1} c_{in} N(\mathbf{x} | \mathbf{\mu}_n, \mathbf{\Sigma}_n) - \sum_{n=0}^{N-1} c_{2n} N(\mathbf{x} | \mathbf{\mu}_{2n}, \mathbf{\Sigma}_{2n}) \right] \ln c_{in} \cdot \sum_{n=0}^{N-1} c_{2n} N(\mathbf{x} | \mathbf{\mu}_{2n}, \mathbf{\Sigma}_{2n}) \quad , (5)
\]

\[
D = \sum_{n=0}^{N-1} c_{in} N(\mathbf{x} | \mathbf{\mu}_n, \mathbf{\Sigma}_n) \quad \forall x \in R_n \quad , (6)
\]

where \( R_n \) is the region of the \( n \)th mixture component. The probability distribution can be approximated by a single Gaussian in this region, i.e.,

\[
\sum_{n=0}^{N-1} c_{in} N(\mathbf{x} | \mathbf{\mu}_n, \mathbf{\Sigma}_n) = c_{in} N(x | \mathbf{\mu}_n, \mathbf{\Sigma}_n) \quad \forall x \in R_n \quad . (8)
\]

Then, the divergence measure between two distributions can be approximated by

\[
D = \sum_{n=0}^{N-1} c_{in} N(\mathbf{x} | \mathbf{\mu}_n, \mathbf{\Sigma}_n) \ln c_{in} \cdot \sum_{n=0}^{N-1} c_{2n} N(\mathbf{x} | \mathbf{\mu}_{2n}, \mathbf{\Sigma}_{2n}) \quad . (9)
\]

Comparing the above divergence measure with the original definition of divergence in Eq. (1), we find that they have the same form. Eq. (9) can therefore be treated as a divergence of two discrete random variables.

In order to calculate the divergence of two audio segments, the component weights \( c_{in} \) in the CCGMM must be estimated first. By using the EM algorithm, they can be estimated by the following iterative formula

\[
\tau_{in} = \frac{\sum N(\mathbf{x}_t | \mathbf{\mu}_n, \mathbf{\Sigma}_n)}{\sum N(\mathbf{x}_t | \mathbf{\mu}_m, \mathbf{\Sigma}_m)} \quad , \quad n = 1, \ldots, N \quad , (10)
\]

where \( t \) is the time index over each audio segment. And, by properly choosing the initial parameters, Eq. (10) will converge rapidly.

Lastly, a simple detection algorithm is used to find the speaker-and-environment change points from the divergence measure curve. If we assume that the mixture weights change linearly when the analysis window crossing a change point, then the divergence can be approximated using a quadratic form. We therefore smooth out the divergence measure curve by convolving it with a weighting function, and then picked up all local maxima to be assigned as change points; i.e.,

Assign \( i \) as a change point if

\[
D'(i) > D'(i-1), \quad D'(i) > D'(i+1), \quad \text{and} \quad D'(i) > \text{Threshold};
\]

where

\[
D'(i) = \sum_{j=i-W}^{i+W} \left[ \left( 1 - \frac{|j-i|^2}{W} \right) D(j) \right] , (11)
\]

and \( W \) equals the width of analysis window divided by the time shift of candidate point.

3. Database and Experiments

3.1. Database

A television broadcast news database was used in the following experiments to evaluate the performance of the proposed method. It was recorded by the Public Television Service Foundation of Taiwan and is referred to as the Public Television Service News Database (PTSNDB). Each recording in the database consisted of a broadcast news episode of 60 minutes. A digital audio recorder (DAT) was used to record the database from the broadcasting machine. The signals were transformed to the form of 16-bit data with 16-kHz sampling rate. In the record, there included opening music, news report, weather report, and advertisement. And, the speakers included the studio anchors, field reporters, interviewees, weather anchors. The background conditions included clean, background music, noise and speech. The corpus was segmented, labeled and transcribed manually using the “Transcriber” developed by LDC. The transcripts were in BIG5-encoded form, with Standard Generalized Markup Language (SGML) tagging to annotate acoustic conditions, background conditions, story boundaries, speaker turn boundaries and audible acoustic events, such as hesitations, repetitions, vocal non-speech events, external noise, etc.

Both orthographic transcription level and acoustic background level markers [5] were extracted from the transcription information as correct answer of the following speaker-and-environment change detection experiments.

Four hours data in PTSNDB were used as the test data in our experiments. Some statistics of environment conditions were shown in Table 1. From Table 1, we can find that there are many diverse audio sources and signal conditions in PTSNDB.

<table>
<thead>
<tr>
<th>Signal Conditions</th>
<th>Percent (in time)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech only</td>
<td>36.0%</td>
</tr>
<tr>
<td>Speech with background music</td>
<td>36.9%</td>
</tr>
<tr>
<td>advertisements</td>
<td>10.0%</td>
</tr>
</tbody>
</table>

Table 1. Statistics of 4 hours PTSNDB database.
3.2. Experiments and Results

In our experiments, all audio signals were pre-emphasis by \( 1 - 0.97z^{-1} \) and segmented into 30-ms frames with 10-ms frame shift. Twelve mel-frequency cepstral coefficients (MFCCs) were then extracted for each frame and taken as feature vectors. We first used one hour data to train a GMM model and took all its mixture components as the common components of CCGMM. Then, divergence measures were computed for candidate points which were equally spaced every 0.5 second over the whole 1-hour data set. For each candidate point, the divergence measure was calculated using Eq. (9) to measure the dissimilarity of the distributions of feature vectors in the two windows beside it. Fig. 1 shows the schematic diagram of the divergence measure calculation for speaker-and-environment change detection.

![Figure 1. The schematic diagram of the divergence measure calculation for speaker-and-environment change detection.](image)

To show the effectiveness of the proposed CCGMM modeling method, an example is displayed in Fig. 2. In this example, the window length is 3 sec (300 frames) and the number of mixtures used in CCGMM is 256. As shown in Fig. 2(a), there are 6 transcription level changes and 3 background condition changes in 50 sec audio signal. In Fig. 2(c), CCGMM weights of four pairs of consecutive windows are displayed. Weights of 10 common components corresponding to the largest weights of the second window are shown in gray level. It can be found from the figure that weights of the second and fourth window-pairs, which correspond to change points, are very different to each other. Fig. 2(d) shows the divergence measure curve and its smooth version. As shown in Fig. 2(d), all local maxima of the \( D'(i) \) curve coincide well with the speaker-and-environment change points.

The proposed method was then further tested using the four-hour data of PTSND shown in Table 1. Various testing conditions with three different numbers of mixture components in CCGMM were examined. They are 256, 128 and 64. With the use of 3-second analysis window, a change point was considered missing if there were no change points detected within a 3-second window centered on the true change point. Fig. 3 shows the false alarm rate vs. miss detection rate curves of these three testing conditions with different threshold values. A miss detection rate (MDR) of 21.9% with 16.0% false alarm rate (FAR) was achieved. As the number of mixture components reduced from 256 to 64, the performance degraded by about 1% in both MDR and FAR.

![Figure 3. The FAR-MDR curves of the proposed speaker-and-environment change detection scheme.](image)

A variation of the proposed CCGMM-based method was then examined. Since the mixture components corresponding to silence had no discrimination capability, we removed them from the CCGMM. By detecting all mixture components \( N(x | \mu_s, \Sigma_s), \forall s \in S \) corresponding to silence, we eliminated the silence effect by calculating new CCGMM weight vector using

\[
C' = \frac{1}{\sum_{i \in S} c_i} \left[ \frac{c_{i,d=0,\ldots,N-1}}{c_{i,d=0,\ldots,N-1}} \right].
\]  

(11)

Nine common components corresponding to silence were moved. A MDR of 21.8% with 16.0% FAR was achieved. The improvement is negligible. This may owing that the duration of silence within speech segment is generally very short.

Then, all false alarms were carefully checked. Table 2 lists the statistics of the analysis. It can be found from Table 2 that the FAR in pure speech is small.

<table>
<thead>
<tr>
<th>Types of conditions</th>
<th>Percent of all false alarms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pure speech</td>
<td>9%</td>
</tr>
<tr>
<td>Speech with music/others + weather reports + advertisements</td>
<td>69%</td>
</tr>
<tr>
<td>Pure music</td>
<td>22%</td>
</tr>
</tbody>
</table>

Lastly, we compared the proposed method with a BIC-based method [5]. In [5], the task of speaker-and-environment change detection based on the BIC-based measure was done on the same PTSND database. Experimental results of the BIC-based method for 3 testing conditions with the audio segments of advertisement and weather report are shown in Table 3. These 3 testing conditions considered change points in different detail. Performances of the proposed CCGMM-
based method for the same 3 conditions but without removing the audio segments of advertisement and weather report are also shown in Table 3. Actually, as shown in Table 2, the FARs in segments of advertisement and weather report were much higher than those for other segments. So the proposed method outperformed the BIC-based method. For the first two testing conditions in Table 3, about 30% reductions in MDR were achieved under the same FAR. If only the speaker turns were concerned, 60% MDR reduction was achieved under the same FAR.

Table 3. Performance comparison for BIC [5] and CCGMM methods in different testing conditions.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>both transcription and background</td>
<td>MDR : 32.3%</td>
<td>MDR : 21.9%</td>
</tr>
<tr>
<td>levels</td>
<td>FAR : 14.5%</td>
<td>FAR : 16.0%</td>
</tr>
<tr>
<td>Transcription level</td>
<td>MDR : 27.0%</td>
<td>MDR : 19.5%</td>
</tr>
<tr>
<td>levels</td>
<td>FAR : 15.6%</td>
<td>FAR : 14.5%</td>
</tr>
<tr>
<td>Speaker turns only</td>
<td>MDR : 22.3%</td>
<td>MDR : 9.8%</td>
</tr>
<tr>
<td>levels</td>
<td>FAR : 40.2%</td>
<td>FAR : 35.5%</td>
</tr>
</tbody>
</table>

4. Conclusions

In this paper, a GMM with common mixture components was proposed to model the audio signal in broadcast news for speaker-and-environment change detection. Experimental results confirmed that the proposed CCGMM-based method outperformed the conventional BIC-based method due to its more precise modeling of audio signal.

5. Acknowledgements

This work was supported by the National Science Council, Taiwan, ROC, under the project with contract NSC 92-2213-E-009-046 and Ministry of Education under the project contract A-93-E-FA06-4-4.

6. References


