THE SIGNAL CHANGE-POINT DETECTION USING THE HIGH-ORDER STATISTICS OF LOG-LIKELIHOOD DIFFERENCE FUNCTIONS

Yih-Ru Wang

National Chiao Tung University,
1001 Ta Hseuh Road, Hsinchu, Taiwan, ROC
Email : yrwang@mail.nctu.edu.tw

ABSTRACT

In this paper, a supervised neural network based signal change-point detector is proposed. The proposed detector uses some high order statistics of log-likelihood difference functions as the input features in order to improve the detection performance. These high order statistics can be easily calculated from the CCGMM coefficients of signals. Performance of the proposed signal change-point detector was examined by using a database of five-hour TV broadcast news. Experimental results showed that the Equal Error Rate (EER) was improved from 16.6% achieved by the baseline method using the CCGMM-based divergence measure to 14.4% by the proposed method.

Index Terms —*Acoustic signal detection, Speech processing*

1. INTRODUCTION

The segmentation of audio signals is an important technology because large amounts of information are delivered every day through audio signals such as broadcast news. The problem is not trivial because an audio signal may contain rich contents generated by various speakers with diverse signal conditions. A good segmentation scheme is useful for further processing, like categorization of broadcast news, speaker diarization or text summarization system.

Many methods were proposed in the past to solve the audio signal segmentation problem. They can be generally categorized into two classes. One adopts the feature-based approach to exploit more effective features for signal segmentation and another uses the metric-based approach to find more efficient measure to describe the similarity/dissimilarity of two signals. In the feature-based studies, lots of distinctive features, such as high ZCR ratio (HZCRR), low short time energy ratio (LSTER), spectrum flux (SF) [1], variance of the spectrum flux (VSF) and variance of the zero-crossing rate (VZCR) [2], were proposed in order to segment complicated audio signals such as broadcast news. The metric-based approach tries to find a good measure which can indicate the statistical similarity/dissimilarity between two audio segments beside a candidate change-point. Many similarity measures were proposed in the past. They included the symmetric Kullback Leibler distance (KL2) [3], divergence shape distance [4] and Bayesian information criterion (BIC) [2,5]. They were all derived from the Jeffrey divergence measure [6] defined as the expectation value of log-likelihood difference between two signals. In order to result in a simple close-form of the Jeffrey divergence measure, the signal was usually assumed Gaussian distributed. In our previous study [7], a more precise signal modeling method, the common component Gaussian mixture model (CCGMM), was used to model audio signals. By using the CCGMM coefficients, the KL2 distance can be simplified and expressed in a discrete form. Moreover, its performance has been shown to be better than other similarity measures such as Bayesian Information Criterion (BIC). Basically, KL2 distance is a function of the first order statistics of the log-likelihood differences between two signals. If some high-order statistics of the log-likelihood differences can be used, the performance of change-point detection should be improved. This motivated us in this study to firstly model mean, variance and skewness of the log-likelihood difference function of two audio signal segments as CCGMM, then take those parameters as detection features, and lastly use an unsupervised neural network-based signal change-point detector to perform audio signal segmentation.

The paper is organized as follows. Section 2 describes the CCGMM-based divergence measure and the proposed multi-layer perceptron (MLP) based signal change-point detector using high order statistics of the log-likelihood difference functions of two audio signal segments. Section 3 discusses the experimental results of using a 5-hour television broadcast news database. Some conclusions are given in the last section.

2. CCGMM-BASED DIVERGENCE MEASURE AND DETECTOR USING HIGH-ORDER STATISTICS OF LOG-LIKILIHOOD DIFFERENCES

2.1. KL2 distance using CCGMM coefficients

The KL2 distance is widely used to measure the similarity of two random variables. It is derived from the average discriminating information between the two random signals and can be expressed by

$$D_2(p_1, p_2) = \int \left[p_1(\mathbf{O}) - p_2(\mathbf{O})\right] \ln \frac{p_1(\mathbf{O})}{p_2(\mathbf{O})} d\mathbf{O}\,, \tag{1}$$

where $p_1(\mathbf{O})$ and $p_2(\mathbf{O})$ are the probability distributions of the two signals. In our previous study, the distributions of two audio segments were represented by CCGMMs, which are mixture Gaussian densities with common mixture components, i.e.,

$$p_n(\mathbf{O}|\lambda_n) = \sum_{i=0}^{M-1} c_{in} N(\mathbf{O}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \quad \forall n = 1,2 \tag{2}$$

where $\{(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i); i = 0, \cdots, M-1\}$ are the means and covariance matrices of the common mixture components of CCGMMs, $\{c_{ni}; i = 0, \cdots, M-1, n = 1, 2\}$ are the mixture weights of the two distributions, and $\lambda_n = \{(c_{ni}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i); i = 0, \cdots, M-1\}$, for $n = 1, 2$, are the parameter sets of the two models.

Then, a divergence measure between the CCGMM distributions of two adjacent audio segments, $\mathbf{O}^S$ and $\mathbf{O}^{S'}$, can be approximated by [7]

$$D_2(\mathbf{O}^S, \mathbf{O}^{S'}) \approx \sum_i (c_{is} - c_{is'}) \ln \frac{c_{is}}{c_{is'}}. \tag{3}$$

Comparing the above divergence measure with the original definition of KL2 divergence shown in Eq. (1), we find that they have the same form. The CCGMM-based divergence measure can therefore be treated as a divergence measure of two discrete random variables. So the divergence of two complicated audio signal segments can be evaluated more easily by using CCGMMs.

In our previous study, a global diagonal covariance matrix $\boldsymbol{\Sigma}$ was used for all mixture components in order to achieve better approximation in the calculation of Eq. (3), The CCGMM of an audio segment then becomes

$$p(\mathbf{O}^S \mid \lambda) = \sum_i p(\mathbf{O}^S, i \mid \lambda) = \sum_i c_{is} N(\mathbf{O}^S; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}). \tag{4}$$

The above CCGMM with global covariance matrix can be regarded as a model of using a set of Parzen windows with Gaussian kernels to estimate the distribution of a signal source. The mixture coefficients of a CCGMM can hence be used to efficiently encode the data samples. Moreover, the divergence of signal sources can be transformed into the divergence of CCGMM coefficients.

2.2. MLP-based signal change-point detector using high-order statistics of log-likelihood difference

Basically, the KL2 divergence measure is just an average of two Kullback Leibler divergence measures:

$$D_2(p_1, p_2) = D_1(p_1 \mid p_2) + D_1(p_2 \mid p_1). \tag{5}$$

The Kullback Leibler divergence measure, $D_1(\ \mid\ )$, is in fact the expectation of log-likelihood difference function defined by

$$D_1(p_1 \mid p_2) = E_{p_1}\left[\log\left(p_1(\mathbf{O})\right) - \log\left(p_2(\mathbf{O})\right)\right]; \tag{6}$$

where p_1 and p_2 are the probability density function of two adjacent audio signal segments.

But, we guess that the following measure maybe more suitable for describing the similarity of two signals:

$$\begin{aligned} D_t(p_1 \mid p_2) &= P\left(\left|\log\left(p_1(\mathbf{O})\right) - \log\left(p_2(\mathbf{O})\right)\right| > t; p(\mathbf{O}) = p_1\right) \\ &= P\left(|d_1| > t\right); \end{aligned} \tag{7}$$

where t is a threshold which can be treated as a tolerance of similarity. The new similarity measure depends on the distributions of the log-likelihood differences rather than only the mean. A new symmetric similarity measure can hence be defined as

$$\begin{aligned} D_{t2}(p_1, p_2) &= \left(D_t(p_1|p_2) + D_t(p_2|p_1)\right)/2 \\ &= \left(P\left(|d_1| > t\right) + P\left(|d_2| > t\right)\right)/2 \end{aligned} \tag{8}$$

It is noted that the new symmetric similarity measure is in the range of [0, 1]. When using the new symmetric similarity measure we need to estimate the distributions of the log-likelihood difference functions, d_1 and d_2, of the two audio signal segments to be modeled. But this is a difficult task. The problem is simply solved in this study by finding some high-order statistics, such as variance and skewness, of the log-likelihood difference functions using the CCGMM coefficients of the two signal segments. Moreover, a supervised neural network based signal change-point detector is also suggested to use the mean, variance and skewness of the log-likelihood difference functions of two audio signal segments as the input features to perform audio signal segmentation. We discuss them in more detail as follows.

First, the means of log-likelihood difference functions can be expressed by

$$\begin{aligned} \mu_d^1 &= E_1\left[d_1(\mathbf{O})\right] = \int p_1(\mathbf{O}) \log\left(\frac{p_1(\mathbf{O})}{p_2(\mathbf{O})}\right) d\mathbf{X} \cong \sum_{i=0}^{M-1} c_{1i} \log\left(\frac{c_{1i}}{c_{2i}}\right) \\ \mu_d^2 &= E_2\left[d_2(\mathbf{O})\right] = \int p_2(\mathbf{O}) \log\left(\frac{p_2(\mathbf{O})}{p_1(\mathbf{O})}\right) d\mathbf{X} \cong \sum_{i=0}^{M-1} c_{2i} \log\left(\frac{c_{2i}}{c_{1i}}\right) \end{aligned} \tag{9}$$

By using the same approximation derived in Eq. (3) [7], the variance of log-likelihood differences can be represented by

$$\begin{aligned} Var_1\left(d_1\right) &\cong \sum_{i=0}^{M-1} c_{1i} \left(\log\left(\frac{c_{1i}}{c_{2i}}\right)\right)^2 - \left(\mu_d^1\right)^2 \\ Var_2\left(d_2\right) &\cong \sum_{i=0}^{M-1} c_{2i} \left(\log\left(\frac{c_{2i}}{c_{1i}}\right)\right)^2 - \left(\mu_d^2\right)^2 \end{aligned}. \tag{10}$$

Thus, the standard deviations of the log-likelihood differences can be expressed by

$$\sigma_d^1 = \sqrt{Var_1\left(d_1\right)};\ \ \sigma_d^2 = \sqrt{Var_2\left(d_2\right)}; \tag{11}$$

Since the log-likelihood differences are not necessarily symmetric, their skewnesses are also calculated by

$$\begin{aligned} skew_d^1 &= \left(E_1\left[\left(d_1 - \mu_d^1\right)^3\right]\left(\sigma_d^1\right)^{-3}\right)^{1/3} \\ &= \left(\sum_{i=0}^{M-1} c_{1i}\left(\log\left(\frac{c_{1i}}{c_{2i}}\right) - \mu_d^1\right)^3\right)^{1/3} \left(\sigma_d^1\right)^{-1} \\ skew_d^2 &= \left(\sum_{i=0}^{M-1} c_{2i}\left(\log\left(\frac{c_{2i}}{c_{1i}}\right) - \mu_d^2\right)^3\right)^{1/3} \left(\sigma_d^2\right)^{-1}. \end{aligned} \tag{12}$$

After finding these high-order statistics of the log-likelihood difference functions between two signal segments, a multi-layer perceptron (MLP) with one hidden layer is used to learn the

similarity measures from the training data in which all signal change-points are properly annotated in advance. The input features of the MLP signal change-point detector are the six parameters found by Eqs. (9), (11) and (12). The similarity measure is now defined in [0, 1]. The desired output of the MLP detector is set to 1 for each signal change-point. Since the window widths of the two audio signal segments, $\mathbf{O}^S$ and $\mathbf{O}^{S'}$, are usually larger than the time shift of candidate signal change points, the output of the MLP detector will be highly correlated with those of neighboring candidate signal change points. If we assume that the CCGMM coefficients change linearly when the analysis window shift across a true signal change-point, the target function of the MLP detector, at time index *k,* can be set as

$$T_k = \begin{cases} 1 - \underset{B_i}{MIN}|k - B_i|/N & ; \underset{B_i}{MIN}|k - B_i| < N \\ 0 & ;\text{otherwise} \end{cases}, \qquad (13)$$

where B_i is the time index of a signal condition change-point in the training data, and the window-width of an audio signal segment is *N* times larger than the time shift of candidate signal change points.

Lastly, a simple detection algorithm is used to find the signal change-points from the output of the MLP detector by picking up all local maxima and comparing with a threshold value.

3. EXPERIMENT RESULTS

A television broadcast news database was used to check the effectiveness of the proposed signal change-point detector using high-order statistics of the log-likelihood difference functions.

3.1 Database

The television broadcast news database was recorded by the Public Television Service Foundation of Taiwan and is referred to as the Public Television Service News Database (MATBN) [8]. Each recording in the database consisted of a broadcast news episode of 60 minutes. In the recording, there included opening music, news report, weather report, and advertisement. The speakers involved included the studio anchors, field reporters, interviewees, and 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. All transcripts were in BIG5-encoded form with Standard Generalized Markup Language (SGML) tagging to annotate acoustics 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 were extracted from the transcription information as the correct answer of the following signal change-point detection experiments.

One hour data in MATBN was used as the training data set in the following experiments. There were in total 428 signal change-points in the training data set. Besides, another four-hour data in MATBN, with 1789 change-points, were used as the test data set. Some statistics of the environment conditions were shown in Table 1. From Table 1, we find that there are many diverse audio sources and signal conditions in the MATBN database. There are about 50 speakers found in each one-hour recording. Moreover, the speaker-and-environment signal conditions changed rapidly. About 50% of them changed within 5 seconds. So, the signal change-point detection task is a difficult one.

Table 1: Statistics of the 4-hour MATBN database.

Signal Conditions	Percent (in time)
Speech only	36.0%
Speech with background music	42.5%
advertisements	10.0%
Music or background sound only	8.9%
Silence	2.6%

3.2 Experimental Results

In our experiments, all broadcast news recordings were pre-emphasized by $1 - 0.97z^{-1}$ and segmented into 30-ms frames with a 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 recording data to train the CCGMM model with all mixture components using the same covariance matrix. Another four-hour recording data was used as the test database. The window length used to find the CCGMM coefficients of audio signal was set to 3 seconds and the number of mixtures used in CCGMM was 256. Then, divergence measures were computed for all candidate change-points which were equally spaced every 0.5 second over the test database. In order to reduce the computation time, the CCGMM coefficients were computed for each 0.5 second sub-window. Then the six sub-windows' CCGMM coefficients in a 3-second analysis window were then averaged. 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 in the following experiments.

First, the baseline scheme of using CCGMM-based divergence measure to detect the signal change-points was tested. To show the effectiveness of the CCGMM modeling method, an example is displayed in Fig. 1. As shown in Fig. 1(a), there are 6 transcription level changes and 3 background condition changes in the 50-sec audio signal segment. In Fig. 1(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 from each other. The false alarm rate (FAR) vs. miss detection rate (MDR) curve of the test data, with different threshold values for divergence measure, is depicted in Fig. 2. An equal error rate (EER) of 16.6% was achieved.

Then, the proposed MLP-based detector using high-order statistics of log-likelihood difference, as shown in Fig. 1(d), was tested. The number of neurons in the hidden layer was empirically set to 15. The results were also shown in Fig. 2. An EER of 14.4% was achieved. About 12% EER reduction can be achieved by using the proposed method.

By further error analysis we found that, for the proposed MLP-based detector operating in the EER condition, the missing detection of speech beginning and ending points was only 9.5%, which was much lower than 14.4% (i.e., EER). On the other hand, we also found that only about 12% of data belonged to non-speech, but about 50% of FA happened at the non-speech part with

background music and sound. They may result from the rapid changing of the signal's characteristics in those segments.

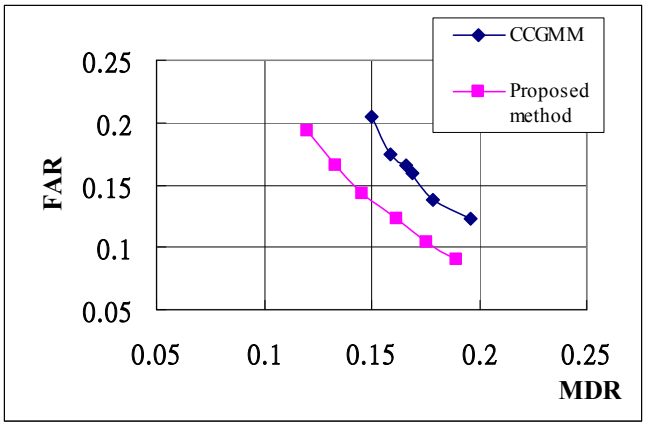

Fig. 2. The FAR-MDR curves of the CCGMM signal change-point detector and MLP-based detector using high-order statistics of log-likelihood differences.

4. CONCLUSIONS

In this paper, some high order statistics, including mean, variance and skewness, of the log-likelihood differences of two audio signal segments were used in audio signal segmentation. These high order statistics features can be easily found by using the CCGMM coefficients of signals. A supervised neural network based detector was suggested to detect signal-change points by using these input features. An ERR reduction of 12% was achieved as tested on a four-hour TV broadcast news database. This result confirmed the effectiveness of the proposed approach.

Acknowledgment

This work was supported by the National Science Council, Taiwan, ROC, under the project with contract NSC 96-2221-E-009 -041.

5. REFERENCES

[1] E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multi feature Music/Speech Discriminator", Proc. ICASSP 97, vol II, pp 1331-1334. IEEE, April 1997.

[2] Rongqing Huang, and John H. L. Hansen, "Advances in Unsupervised Audio Classification and Segmentation for the Broadcast News and NGSW Corpora", IEEE Transactions on audio, speech, and language processing, vol. 14, no. 3, may 2006, pp. 907-919.

[3] M.A. Siegler, U. Jain, B. Raj, R. M. Stern, "Automatic Segmentation, Classification of Broadcast News Audio", *Proc. DARPA speech recognition workshop*, pp. 97-99, 1997.

[4] L. Lu, H.-J. Zhang, "Speaker change detection and tracking in real-time news broadcasting Analysis", *Proc. of ACM Multimedia 2002*, pp. 602-610.

[5] Scott Shaobing Chen, P.S. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion", *Proc. DARPA speech recognition workshop*, 1998.

[6] H. Jeffreys, "An Invariant Form for the Prior Probability in Estimation Problems", *Proc. Roy. Soc. Lon., Ser. A*, no. 186, 453-461, 1946.

[7] Yih-Ru Wang and Chi-Han Huang, "Speaker-and-environment Change Detection in Broadcast News using the Common Component GMM-based Divergence Measure", Proc. of ICSLP 2004, Jeju island, pp. 1069-1072, Oct. 2004.

[8] Hsin-Min Wang, Shi-Sian Cheng and Yong-Cheng Chen, "The SoVideo Mandarin Chinese News Retrieval System", *Int. Journal of Speech Technology*, Vol. 7, pp 189-202, 2004.

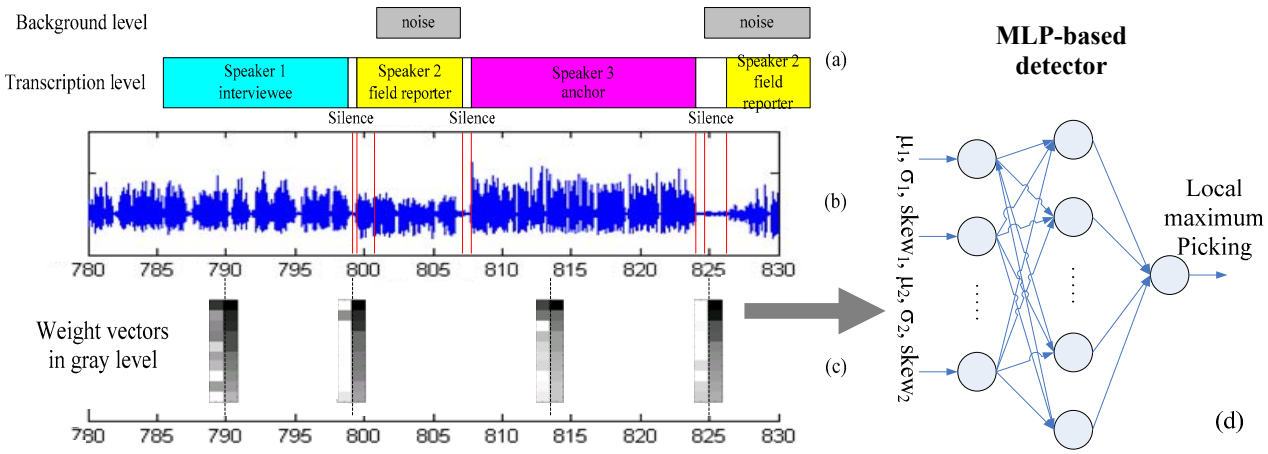

Fig. 1. An example of the proposed CCGMM-based method of signal change-point detection in broadcast news signals: (a) transcription information, (b) waveform (vertical lines indicate marks of change points), (c) weights of 10 common components corresponding to the largest weights of the second window shown in gray level, (d) the proposed MLP-based signal change-point detector using high-order statistics of log-likelihood difference functions.