ROBUST UNSUPERVISED SPEAKER TURN DETECTION

Assefa Kassa Teshome
Department of Electrical and Computer Engineering, EiT-Mekelle, Mekelle University
Mekelle, Ethiopia
Assefa_kt@yahoo.com

and

C.S. Ramalingam
Department of Electrical Engineering, IIT-Madras
Chennai 600-036, India
csr@iitm.ac.in

ABSTRACT

In this paper we address an aspect of speaker recognition task, viz. unsupervised speaker turn detection. A metric based approach with two-pass criteria is proposed for this task. A GMM-based modified Log Likelihood Ratio metric is used in the first pass; Bayesian Information Criterion (BIC) metric is used in the second pass to verify or discard the speaker turn points hypothesized in the first pass. We consider two cases: long speaker turn segments (> 2 sec.) and short speaker turn segments (< 2 sec.). We have evaluated our algorithm using TIMIT speech files. Our precision results range from 85% to 93%, recall ranges from 75% to 78%, and the F-ratio is in the range 80—85%. These results are better than what has been reported in the literature so far.

Keywords: Bayesian Information Criterion (BIC), Log Likelihood Ratio (LLR), speaker change detection, Speaker Turn Indexing, Speaker Turn Detection.

1. INTRODUCTION

Automatic speech and speaker recognition are beginning to play an increasingly important role in our daily lives. The need for such systems is felt in transcription, data mining, etc. Distinguishing the change in speaker when many persons are conversing is needed in applications such as videoconferencing, where we wish to pan the camera automatically towards the current speaker. Speaker turn separation also finds application in data mining.

There are two ways in which the problem of speaker turn separation can be addressed, i.e., metric- and model-based approaches. The model-based approach requires *a priori* knowledge of number of speakers, which is not practical in many applications. Hence, the metric-based approach is often the preferred choice. Most metric-based approaches formulate the problem as follows: decide whether or not a speaker change point exists at time 't' based on processing data in neighboring windows of relatively small size around time `t', as shown in Fig. 1. The content of these windows are usually sequences of feature vectors extracted from the audio signal. In Fig. 1, these sequences are denoted by $X = \{x_1, x_2, x_3, \ldots, x_{Nx}\}$ and $Y = \{y_1, y_2, \ldots, y_{Ny}\}$, where N_x, and N_y are the number of data points in the two windows. Let Z denote the concatenation of the contents of the two windows having N_z data points. The contents of these two windows are compared using a dissimilarity function. Local extrema of this dissimilarity function are compared to a threshold, yielding possible speaker-change points.

Jitendra et al. [1] used a modified log likelihood ratio (mod LLR), metric to achieve reasonably robust speaker separation. Perrine et al. [2] used a two-pass, where in the first pass speaker turn points are hypothesized using a generalized likelihood ratio distance. The likelihoods are computed based on single Gaussian models used for representing the current analysis window. The second pass criterion uses Bayesian Information Criterion (BIC) to validate or discard the potential speaker change candidates by the first pass criterion. Kemp at al. [3] used energy based speaker segmentation technique. Jørgensen et al. [4] used the Kullback–Leibler metric along with a VQ model. Adami et al. [5] proposed a two-speaker separation algorithm based on the crossing points of GLR distance curves plotted with respect to the estimated models of each speakers. Segmentation algorithms based on distance between two consecutive parts of speech signal have been investigated in [6]--[8]. A segmentation algorithm based on BIC is presented in [9] but require long speech segments.

Our proposed method combines the advantageous features of the various segmentation techniques to improve performance. Our approach is similar to the one used in [2]. But, unlike in [2], we have used a modified Log Likelihood Ratio instead of GLR and a GMM model instead of single Gaussian approximation for each analysis window. These have led to much better performance.

Our paper is organized as follows: Section 2 summarizes the likelihood based metric dissimilarity measures. Section 3 presents the proposed criterion. Section 4 explains the experimental setup. Section 5 shows the results. Section 6 shows the conclusions.

2. METRIC DISSIMILARITY MEASURES

In general, all metric-based approaches use hypothesis testing to decide whether or not there is a speaker change. H0 is the null hypothesis, i.e., no speaker change, whereas H1 is the alternative hypothesis, i.e., speaker change exists.

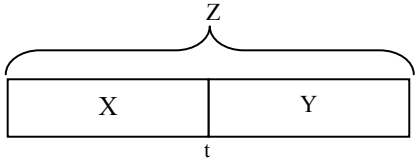

Figure 1. *Analysis windows for metric dissimilarity*

$$H0 : (z_1, ..., z_{N_z}) \sim \mathrm{N}(\mu_z, \Sigma_z)$$
$$H1 : (x_1, ..., x_{N_x}) \sim \mathrm{N}(\mu_x, \Sigma_x)$$
$$and\ (y_1, ..., y_{N_y}) \sim \mathrm{N}(\mu_y, \Sigma_y)$$

2.1 Generalized Likelihood Ratio (GLR) Distance

The Generalized Likelihood Ratio (GLR) tests the H0 and H1 hypotheses based on the ratio of the likelihoods in favor of each hypothesis. The likelihood L_0 for H0 hypothesis is the probability that all the acoustic features are in Z given the model λ_z.

$$L_0 = \prod_{i=1}^{N} p(Z_i / \lambda_z) \quad (1)$$

$$\Rightarrow L_0 = \prod_{i=1}^{N_x} p(x_i / \lambda_z) \prod_{i=1}^{N_y} p(y_i / \lambda_z) \quad (2)$$

The alternative hypothesis H1 will be true when X and Y come from different speakers and hence from different GMM models. That is, X is from λ_x whereas Y is from λ_y. As a result likelihood L_1 for H1 is given by:

$$L_1 = \prod_{i=1}^{N_x} p(x_i / \lambda_x) \prod_{i=1}^{N_y} p(y_i / \lambda_y) \quad (3)$$

The ratio of the likelihoods of the two hypotheses is then:

$$L = \frac{L_0}{L_1} = \frac{L(Z / \lambda_z)}{L(X / \lambda_x) L(Y / \lambda_y)} \quad (4)$$

$$GLR(t) = -\log(L) \quad (5)$$

where λ_x, λ_y, and λ_z are the GMM models for X, Y, and Z respectively.

Since the log of the ratio is taken, GLR is also known as the Log Likelihood Ratio (LLR) in most literature.
This computation is to be carried for each analysis window. Duration of X and Y is typically 1 sec. and a new Z is shifted by the desired resolution of finding potential change points. Finally, the candidate speaker change points will be the locations of the local maxima of the GLR distance curve compared to a certain threshold. The threshold is entirely experimental. The difficulty of the problem lies in finding global optimum for the threshold.

2.2 Bayesian Information Criteria (BIC)

The BIC procedure is entirely based on the log likelihood ratio except that the BIC is a likelihood criterion penalized by the model complexity. Considering Fig. 1 and L(Z/M) the likelihood of Z for the model M, the BIC value is determined by

$$BIC(M) = \log L(Z / M) - \lambda \frac{m}{2} \log N_z \quad (6)$$

where *m* is the number of parameters of the model M and λ the penalty factor. We consider the following hypothesis test for speaker change at time t:

Log likelihood values L_o and L_1 are computed using the same set of equations used for GLR Eq. (2) and (3).
Ignoring the variation of the mean, as it is easily biased by the channel condition, the maximum log likelihood ratio between hypothesis H0 (no speaker change) and H1 (speaker change at time t) is then defined by [2]:

$$_{\mathrm{V}}R(t) = \frac{N_z}{2} \log |\Sigma_z| - \frac{N_x}{2} \log |\Sigma_x| - \frac{N_y}{2} \log |\Sigma_y| \quad (7)$$

complete sequence, X, Y, and Z. The variation of the BIC value between the hypotheses is then

$$\Delta BIC(t) = -R(t) + \lambda \frac{1}{2} \Delta k \quad (8)$$

Notice that a penalty term λΔk is added; Δk is to account for the difference of number of parameters and is given by

$$\Delta k = (p + \frac{1}{2} p(p+1)) \log N_z$$

where p is the dimension of the features vector

In the decision, a negative value of ΔBIC(t) indicates that the individual models best fit the data Z than the combined single model, which means that a change of speaker occurred at time t. BIC based procedures have proven to be efficient as stated in most literature, but computationally costly.

2.3 Modified Log Likelihood Distance (Mod LLR)

The hypotheses are formulated in the similar manner as the other two cases. The only difference is that in hypothesis H0, the data Z are modeled with a two-mixture GMM instead of a single Gaussian density. The ML estimates of the parameters of this GMM λ'_z, are calculated using the EM algorithm. The log likelihood L'_0, in this case is calculated as

$$L'_0 = \sum_{i=1}^{N_x} \log p(x_i / \lambda'_z) + \sum_{i=1}^{N_y} \log p(y_i / \lambda'_z) \quad (9)$$

Note that $L'_0 \geq L_0$ eq(2), *s*ince a GMM can always fit the data better (or equally well) compared to a single Gaussian density.
The modified LLR criterion distance d_{LLR} is then simply the log likelihood ratio.

$$d_{LLR} = L_1 - L'_0 \quad (10)$$

All the local maxima of d_{LLR} greater than certain threshold θ are considered to be speaker change points. It can be seen that, in contrast to the standard LLR(or GLR) and BIC techniques, all terms in this criterion are derived directly from the data, and thus the criterion can be expected to be robust to changing data conditions [2]. This criterion has simplicity of the GLR and a robustness of the BIC criterion.

3. PROPOSED METHOD

3.1 The First Pass Criteria

In the first pass criteria, a modified implementation of Modified Log Likelihood Ratio is used. To decide whether or not a speaker change point exists at tme t_i (Fig 2), two neighboring windows of relatively small size are considered around each t_i. The content of these windows are sequence of feature vectors extracted from the silence removed speech signal. If silence is not removed, comparison between silence

and speech would be performed resulting in too many insertion errors.

The whole idea of modified LLR is to create equal number of parameters while modeling the two hypotheses. Most researchers have used a single Gaussian for modeling X and Y separately and a two-mixture GMM for modeling the union window Z.

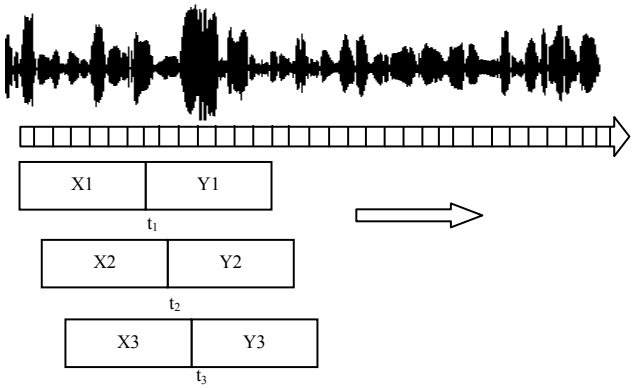

Figure 2. *The 1st pass algorithm*

However, a single Gaussian model is almost always a poorer fit to a speech window of any length compared to a multi-mixture GMM. We therefore modified the implementation by using an n– mixture GMM to model X and Y, and a 2n–mixture GMM to model the union Z. This way we can still maintain balance in the number of parameters that needs to be estimated in both hypotheses while fitting the data with a better likelihood to the corresponding models. The only problem associated here is what n should be. Experimentally we have found out that n=2 gives the best results. The duration of X and Y being 1 sec, using more mixtures will result in "over-modeling". For each X_i and Y_i pairs the likelihoods of H0 and H1 are computed using Eq. (9) (and its corresponding log values). The distance at each analysis stage is computed using Eq. (10). This operation is carried out until end of the speech for every frame shift to get a good resolution of finding the change points. However, the mod LLR distance curve is very noisy. The curve is mean- and variance-normalized and a moving average filter is used repeatedly to low-pass filter unwanted high frequency variations from the curve. At this stage "strong" maxima are selected by comparing the difference between the maximum and its surrounding minima on either side. If the difference is greater than a threshold, then the frame corresponding to that maximum is hypothesized as a possible candidate for speaker change.

3.2 The Second Pass Criterion

This stage is used for either validating or rejecting the candidate points hypothesized in the first pass. The BIC procedure is very effective when there is a larger set of acoustic vectors in each segment. Let $s_0, s_1, s_2, s_3, s_4, s_5, s_6 \ldots$, be the candidates picked up by the first pass criterion. The BIC metric is used in the manner described in Fig. 3 for evaluating whether these points represent genuine turn points.

s_0 is the starting frame of the speech. To validate that s_1 is genuine between s_0 and s_2, the BIC metric in Eq. (8) is calculated by using X to represent acoustic features between s_0 and s_1, and Y between s_1 and s_2. There are two possible cases for the values of ΔBIC(i) computed in this fashion

3.2.1 Case 1: ΔBIC (i) >0

If ΔBIC (i) is positive, the combined model best describes the data than the separate models. As a result the candidate change point 'i' is discarded. If this happens while trying to decide for s_2 the analysis window X will be from s_0 to s_2 and Y will be from s_2 to s_3. This is because of the fact that the BIC procedure qualifies the data from s_0 to s_2 to come from a single speaker. Larger the size of the data more optimal the BIC will be.

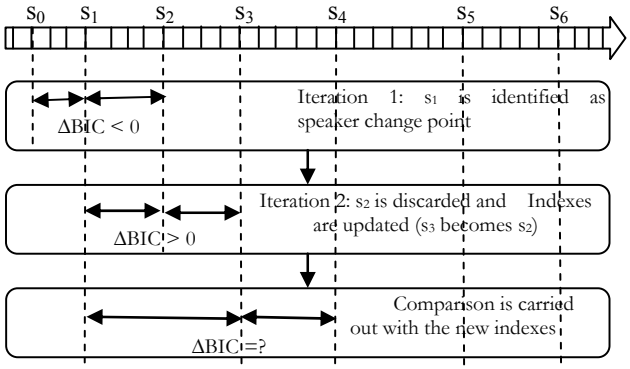

Figure 3. *The 2nd pass algorithm*

3.2.2 Case 2: ΔBIC (i) < 0

In this case, the separate models best describe the data than the combined one. Therefore, the candidate point s_1 is validated as one of the final change points detected. And when trying to decide for s_2 the analysis window X will be from s_1 to s_2 unlike case 1 and Y will be from s_2 to s_3.

This procedure is carried out until all the candidate points by the first pass are evaluated. The second pass criterion has proven to be robust in removing insertion errors without affecting much the genuine speaker change points. This illustrates the power of BIC for this task.

4. EXPERIMENTAL SETUP

To test our approach we use different types of speech data:

- A conversation is artificially created by concatenating 15 sentences of 2 s on average from the TIMIT database (short segments). This file contains different speakers and named TIMIT15ptsShortSeg. These short segments model telephone conversation speech. Analysis windows X and Y are 1 sec.
- A conversation is created by concatenating 8 sentences of 9 s on average from the TIMIT database (long segments). This file models broadcast transmission as it has long segments and is named TIMIT7ptsLongSeg. Analysis windows X and Y are 2 sec.

The speech signal is windowed using 30 ms duration for every 10 ms shift. It is then parameterized with 12 MFCC coefficients. The addition of the Δ-coefficients (first derivatives) does not improve the results and increases the time of computation. For this reason, the Δ-coefficients were not used.

5. RESULTS AND PERFORMANCE EVALUATION

A change detection system has two possible types of error. Type-I error occurs if a true change is not spotted within a certain window (0.5 secs in either sides of the true change, in our case). Type-II error occurs when a detected change does not correspond to a true change in the reference (false alarm) [4]. Type I and II errors are also referred to as precision (PRC) and recall (RCL), respectively, and are defined as

$$PRC = \frac{Number\ of\ correctly\ found\ changes}{Total\ number\ of\ changes\ found} \quad (11)$$

$$RCL = \frac{Number\ of\ correctly\ found\ changes}{Total\ number\ of\ correct\ changes} \quad (12)$$

In order to compare the performance of different systems, the F-measure is often used; it is defined as

$$F = \frac{2.0 \times PRC \times RCL}{PRC + RCL} \quad (13)$$

The F-measure varies from 0 to 1, with a higher F-measure indicating better performance. Comparison with benchmark models is given below.

1.) Results found by Perrine et al. [2]

File	PRC	RCL	F
TIMIT(29pts)	0.759	0.595	0.67
TIMIT(27pts)	0.630	0.607	0.62

Table 1. *Performance evaluation result of BIC based algorithm used by [2]*

File	PRC	RCL	F
TIMIT(29pts)	0.793	0.676	0.73
TIMIT(27pts)	0.815	0.667	0.73

Table 2. *Performance evaluation result of proposed by Perrine et al [2]*

2.) Results found by Jitendra et al. [1] Uses HUB-4 1997 evaluation setup

File	PRC	RCL	F
-	0.68	0.65	0.67

Table 3. *Performance evaluation result of proposed by Jitendra et al [1]*

3.) Proposed Method

File	PRC	RCL	F
TIMIT(15pts) short seg	0.93	0.78	0.85
TIMIT(7pts) long seg	0.857	0.75	0.80
TIMIT(26pts) short seg	0.85	0.78	0.81

Table 4. *Performance evaluation result of our proposed method.*

6. CONCLUSIONS

We have used a metric-based approach, which has the advantage of not requiring a priori knowledge of number of speakers. Specifically, we have used a modified LLR based hypothesis test where the number of parameters used to model the data in the two hypotheses is forced to be the same in the first pass. Thus, the likelihoods in these two hypotheses are directly comparable. We have used a BIC criterion in the second pass to validate or discard the candidate change points at the first pass. We have shown that performance improves significantly compared to that reported in [2] if (1) multi-mixture Gaussian modeling is used instead of single-mixture Gaussian, and (2) modified LLR is used instead of GLR. The usefulness and robustness of this measure was further illustrated with the help of experiments where the proposed criterion achieved better F-measure compared with most of the proposed criteria in previous researches as given in Tables 1, 2, and 3.

Currently our proposed method only works offline. However, many applications require an online speaker indexing systems. This will be part of future work.

7. REFERENCES

[1] Jitendra Ajmera, Iain McCowan, and Hervé Bourlard, "Robust Speaker Change Detection", *IEEE Signal Processing Letters*, Vol. 11, No. 8, August 2004.

[2] Perrine Delacourt et'al "Speaker-based segmentation for audio data indexing", ICMCS, Vol. 2, pp. 959-963, 1999.

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

[4] Kasper Jørgensen et al, "Unsupervised Speaker Change Detection For Roadcast News Segmentation", *Eusipco*, 2006.

[5] André G. Adami et al, "A New Speaker Change Detection Method For Two-Speaker Segmentation", *IEEE*, vol 4, pp 3908-3911, 2002.

[6] M. A. Siegler et al., "Automatic segmentation, classification, and clustering of broadcast news audio", *DARPA speech recognition workshop*, 1997.

[7] C. Montaci´e and M.-J. Caraty, "Sound channel video indexing," *Eurospeech*, pp. 2359–2362, 1997.

[8] H. Beigi and S. Maes, "Speaker, channel and environment change detection," *World congress of automation*, 1998.

[9] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion", *DARPA speech recognition workshop*, 1998.

[10] J. P. Campbell, Jr., "Speaker Recognition: A Tutorial", *Proceedings of the IEEE*, vol. 85, no. 9, pp 1437-1462, September 1997.