A Two Pass Algorithm for Speaker Change Detection

Sunil Kumar Kopparapu, G Sita
TCS Innovation lab - Mumbai,
Tata Consultancy Services Limited,
Yantra Park, Thane (West) - 400 601.
Email: sunilkumar.kopparapu@tcs.com

***Abstract*—Speaker change detection is a necessary first step in several applications. In this paper, we propose an unsupervised two pass algorithm for speaker change detection in conversational speech. Generalized Likelihood Ratio (GLR) metric is used in the first pass to coarsely identify speaker change points and during the second pass, these candidate change points are finely analyzed assuming that the initial part of the conversational audio is from one of the speakers. The final change point detection decision is based on the likelihood probability function computed for the segments between two consecutive candidate change points using the known speaker model. The proposed two pass algorithm has been tested on a question and answer session of a financial audio report of a company and also on an audio track of a movie.**

I. Introduction

Speaker change detection forms a sub-category of a more general problem of audio segmentation to obtain discrete and characteristically similar segments from a given audio stream. Automatic identification of the number of speakers and speaker change detection in conversational speech is a very important problem in voice analytics in general and for speaker identification and verification in particular. In this paper, we consider conversational speech to identify speaker change points with an aim to segregate same speaker related segments. Speaker change detection is a vital component in audio analytics in general and for telephone monitoring systems, speaker tracking and indexing applications in particular. Speaker change detection is a necessary pre processing step in automatic transcription and indexing of broadcast news or movie audio tracks for summarization. Extraction of audio segments corresponding to a specific speaker would also help in pronunciation and accent analysis in a call center scenario. In general segmentation methods can be broadly categorized into two categories, namely, metric based and model based methods. In metric based segmentation methods, distance between two adjacent equal length speech windows is computed for frame-wise[1] feature vectors using metrics like Dynamic Time Warping (DTW), Euclidean or Mahalonabis distance and the speaker change point is detected where the distance between the frames exceeds a *preset* threshold. Most of the distance measuring criteria come from statistical modeling framework. In model based methods, the feature vectors in each of the adjacent windows are generally assumed to follow a certain (usually Gaussian) distribution and the distance between adjacent segments is computed in terms of the dissimilarity of these two distributions using for example Bayes Information Criterion (BIC) or Generalized Likelihood Ratio (GLR) or Kullback-Leibler divergence (KLD) measure[1], [2], [3]. Typically, if there is no speaker change point detected, the analysis window is grown in length so as to obtain more robust statistics, however the window length adds to the computational cost and subsequently hinders real time applications. Improved BIC based approaches were proposed in [4], [5] to speed up the change point detection process. Segmentation based on BIC [6] requires long segments of data for accurate results and also computationally more intensive. In this paper, we assume that the initial part of the audio conversation contains speech from only one speaker, this initial audio segment is used to build models for that speaker. This speaker model information is used in the second pass.

Several speech segmentation algorithms have been proposed in literature. Methods based on BIC are extensively used for this purpose although they require large segments of data to obtain meaningful results. Generally, apriori information is not available on the number of speakers involved in a conversation. Broader scope of speaker change detection scheme is to identify the number of speakers [7] in a conversation, segregating homogeneous parts corresponding to different speakers into different clusters so as to track a particular speaker in a given conversation and to index the various segments. In the current work, we assume two speakers in a given conversation although the method can be extended to multi speaker conversations. The proposed speaker change detection scheme attempts to measure the dissimilarity between two consecutive segments of the parametrized signal to decide if these segments are generated by the same speaker or different speakers. The comparison is carried out in cepstral domain using the MFCC feature vectors of the segments. The feature vectors in each of the segments are assumed to follow Gaussian distribution. We propose a two level segmentation procedure which uses GLR computed locally in the first level to make a coarse segmentation and a refined segmentation decision based on the posterior likelihood probabilities of the identified candidate segments in the second level. We assume that the conversation has only two speakers and also assume simultaneous speech

[1]Framing of speech is commonly used in speech processing. Typically a speech frame is of duration 20 ms.

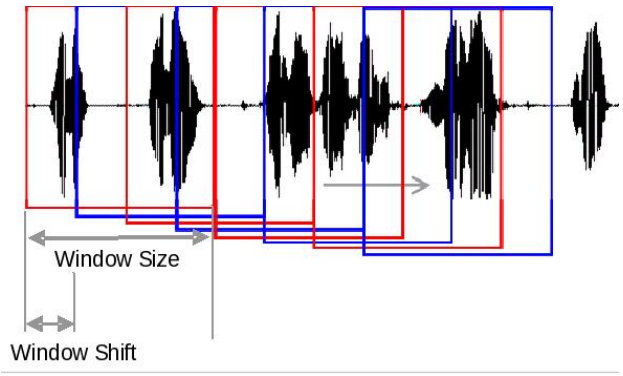

Fig. 1. Adjacent window analysis for speaker change detection

by both speakers does not occur. We first explain the proposed two pass algorithm in detail. The rest of the paper is organized as follows: We explain the GLR criterion and the method of computation of posterior likelihood probabilities in Section II. Then the experimental results obtained using the proposed method along with discussion are presented in Section III followed by conclusions in Section IV.

II. Two Pass Speaker Change Detection

The baseline change point detection technique used in speaker segmentation systems [8] is to divide the entire speech into small overlapping windows of size 25 secs, then compare adjacent windows and decide whether the two adjacent windows of speech have similar characteristics or different, meaning if they originate from the same source or different sources. As shown using Figure 1 the adjacent overlapping windows (represented by a red box and a blue box) are compared to determine if these windows of speech have similar characteristics of different. This decision is done by sliding along the time axis.

For a pair of adjacent windows being compared and found to be originating from different sources, the end point of the first window would be considered the change point. Clearly the window size constrains the detection of short duration changes. Also the localization (closeness of a detected change point to the actual change point) of the detected changes directly depends on the size of window overlap and/or the window shift [9].

A. First Pass: Coarse Detection

We assume that the audio conversation starts with one speaker speaking for at least 10 sec (say *Speaker 1*). This data is used to build a reference model λ_1 for *Speaker 1*. The rest of the audio conversation is analyzed using a sliding window ($\mathcal{W}$) of length 2 sec. A hypothetical segment boundary is assumed at the center of the window with the first part (say $\mathcal{W}_1$) of the window being assumed as a continuation of a speaker and the second part (say $\mathcal{W}_2$) of the window being assumed as generated from the other speaker. We carry out the analysis using the Mel Frequency Cepstral Coefficients (MFCC) feature vectors[2] obtained for the two sub-windows $\mathcal{W}_1$ and $\mathcal{W}_2$ under consideration using speech frames[3] of length 20 ms with a frame overlap of 10 ms. Let there be N frames in each of the two sub-windows $\mathcal{W}_1$ and $\mathcal{W}_2$. We have,

$$\mathcal{W}_1 = \{w_{11}, w_{12}, \cdots, w_{1N}\}$$

and

$$\mathcal{W}_2 = \{w_{21}, w_{22}, \cdots, w_{2N}\}$$

The task is to decide if these two sub-windows belong to the same or different acoustic conditions and hence speakers. Assume that the hypothesis $\mathcal{H}_0$ indicates that the two sub-windows belong to one single multivariate Gaussian process or a single speaker, namely,

$$\mathcal{H}_0 : \mathcal{W}_1, \mathcal{W}_2 \sim N(\mu_{\mathcal{W}}, \Sigma_{\mathcal{W}}) = N_{\mathcal{W}}$$

The hypothesis $\mathcal{H}_1$ indicates that the two segments $(\mathcal{W}_1, \mathcal{W}_2)$ are generated by two different multivariate Gaussian processes or two speakers, namely,

$$\mathcal{H}_1 : \mathcal{W}_1 \sim N(\mu_{\mathcal{W}_1}, \Sigma_{\mathcal{W}_1}) = N_{\mathcal{W}_1} \quad \text{and}$$

$$\mathcal{W}_2 \sim N(\mu_{\mathcal{W}_2}, \Sigma_{\mathcal{W}_2}) = N_{\mathcal{W}_2}$$

where $\mu_{\mathcal{W}}, \mu_{\mathcal{W}_1}$ and $\mu_{\mathcal{W}_2}$ are the sample mean vectors and $\Sigma_{\mathcal{W}}, \Sigma_{\mathcal{W}_1}$, and $\Sigma_{\mathcal{W}_2}$ are the sample covariance matrices of the entire window $\mathcal{W}$ and the two sub-windows $\mathcal{W}_1$, $\mathcal{W}_2$ respectively. The generalized likelihood ratio ($\mathcal{R}_{\mathcal{W}}$) between the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ for the window $\mathcal{W}$ is defined as,

$$\begin{aligned} \mathcal{R}_{\mathcal{W}} &= \log L(\mathcal{W}, N_{\mathcal{W}})) \\ &- (\log L(\mathcal{W}_1, N_{\mathcal{W}_1}) + logL(\mathcal{W}_2, N_{\mathcal{W}_2})) \end{aligned} \tag{1}$$

The GLR ($\mathcal{R}_{\mathcal{W}}$) is computed [3] for a pair of adjacent sub-windows of same size and the analysis window is then shifted by a step length of 0.5 sec along the speech signal and the likelihood ratio for the new window is computed. Negative $\mathcal{R}_{\mathcal{W}}$ indicates that the sub-windows are better represented by $N_{\mathcal{W}_1}$ and $N_{\mathcal{W}_2}$ rather that the whole window $\mathcal{W}$ being represented by $N_{\mathcal{W}}$ meaning $\mathcal{W}$ had a speaker change point. The GLR distances thus computed for all windows for the audio track are computed and a threshold[4] T is used to detect all possible candidate speaker change points. The threshold is so chosen so that we do not miss out on any true change points so that in the second pass analysis there is a chance of the speaker change point being detected.

B. Second Pass: Speaker Change Detection

A second level verification is carried out to narrow down on the most likely speaker change point. The candidate speaker change points detected in the above step are considered sequentially and we try to find the pair of segments that have a high likelihood of being from different speakers. For this we

[2]MFCCs are the most popularly used speech features in most speech and speaker recognition applications

[3]framing the speech signal for the purpose of analysis is a well know in speech literature

[4]chosen empirically through experiments

use the initial part of the data which is assumed to be from *Speaker 1* having a model parameter set represented as, λ_1. The similarity of the sub-window $\mathcal{W}_1 = [w_1, w_2, \cdots, w_N]$ in the neighborhood of a candidate speaker change point detected in the first pass is computed as the log-likelihood

$$S(\mathcal{W}_1|\lambda_1) = \sum_{i=1}^{N} \log(p(w_i|\lambda_1)) \quad (2)$$

The posterior probability that an observed feature vector w_i was generated by the model, λ_1 is given by,

$$p(w_i|\lambda_1) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} exp^{\left\{-\frac{1}{2}\left((w_i-\mu)^T\Sigma^{-1}(w_i-\mu)\right)\right\}} \quad (3)$$

where d is the dimension of the feature vector w_i which is 14 in our experiments[5]. Since true densities of the *Speaker 1* is unknown, they can be approximated by sample mean and variances for computing the speaker model λ_1. The change points for which the adjacent sub-window segments have a large difference in similarity computed from (2) are identified as valid speaker change points. The first valid speaker change point signals the beginning of a second speaker segment. The first speaker model parameters are modified using all the frames up to the detected change point.

III. Experimental Results and Discussion

We used two different data sets to verify the performance of the two pass algorithm to identify speaker change points. The first data set consisted of 30 conversation segments occurring between two speakers in a popular Hindi movie track called MOVIE TRACK. The second data set is from an audio financial report presentation of a company in which at various points different speakers[6] pose questions and seek answers from the company representative, we call this QA TRACK. From QA TRACK, we extracted conversation segments of average length 80 sec to ensure that there is at least one speaker change point in the conversation. We collected 30 such speech conversations from QA TRACK to test our two pass algorithm. We investigated the performance of a single pass algorithm namely using only the GLR criterion (First Pass) and only using likelihood probability (Second Pass) and compared it with the two pass algorithm proposed in this paper. Results are presented for all the three scenarios in Table I. Overlapping speech frames of length 20 ms are generated using Hamming window with a overlap of 10 ms. We use 8 MFCC, 4Δ MFCC and $2\Delta^2$ MFCC a total of 14 Mel Frequency Cepstral Coefficients, not including the average energy or the first MFCC coefficient. It is found that short utterances of length less than 0.5 sec are not recognized as the window shift for change detection is 0.5 sec and short segments do not have enough speaker information leading to erroneous results. It has been observed that with longer window length (< 2 sec), the performance of the two pass algorithm slightly improves as the likelihood statistics are more accurate with more number of frames to characterize a given audio segment. The errors in speaker change detection are categorized as false positives, namely, a speaker change is detected although there is no speaker change at that point and false negative which captures true speaker change points in the data that have been missed by the algorithm. Many of the false negatives are found to be either due to (a) simultaneous speech by both the speakers or (b) speech with less voice activity. The inclusion of a silence detection module resulted in an improved performance of the two pass speaker change detection by 8%. It was also observed that audio signal segments with short coughs and heavy breathing contributed to errors considerably.

Method	Segments Tested	Correct	False +ve	False -ve
One Pass				
Section II-A GLR	MOVIE TRACK 30	19	7	4
	QA TRACK 30	17	4	9
Section II-B MAP	MOVIE TRACK 30	20	4	6
	QA TRACK 30	16	6	8
Two Pass				
GLR-MAP	MOVIE TRACK 30	**23**	4	3
GLR-MAP	QA TRACK 30	**18**	3	4

TABLE I
EXPERIMENTAL RESULTS

It can be observed from Table I that the performance of the two pass algorithm is better than either of the single pass algorithms. The improvement in performance is not only in terms of the increased correct identification of the speaker change points but also in terms of the reduced number of false positives and false negatives.

IV. Conclusion

In this paper we presented a two pass algorithm based on GLR and MAP for speaker change detection for two speaker conversational speech. The method was tested on two different types of audio track, one corresponding to a movie and the other corresponding to a question answer session of a financial meeting. The proposed method performs well on both the movie track and the question answering audio track and exceeds the performance of a single pass scheme. The performance of the speaker change detection is better for the movie track because the audio signal is acquired under controlled conditions and is of high quality while the performance of the two pass algorithm reduced for the question answer audio track which has considerable amount of extraneous noise. We were unable to test the proposed method on a standard database for lack of availability of access to such a database source. Inclusion of a silence removal module before speaker change detection improved the performance by about 8%. We plan to experiment with different speech enhancement modules and extend the method to multiple speaker change detection scenario.

[5] 8 MFCC, 4Δ MFCC and $2\Delta^2$ MFCC
[6] journalists

REFERENCES

[1] Shih-Sian Cheng and Hsin-Min Wang, "A sequential metric-based audio segmentation method via the Bayesian information criterion," in *EUROSPEECH-2003*, 2003, pp. 945–948.

[2] Wei-Ho Tsai, Shih-Sian Cheng, and Hsin-Min Wang, "Automatic speaker clustering using a voice characteristic reference space and maximum purity estimation," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 15, no. 4, pp. 1461–1474, 2007.

[3] B. Narayanaswamy, G. Rashmi, and R. Stern, "Voting for two speaker segmentation," in *Proc. ICSLP*, 2006.

[4] A. Tritschler and R. Gopinath, "Improved speaker segmentation and segments clustering using the bayesian information criterion," in *EUROSPEECH-1999*, 1999.

[5] B. W. Zhou and John H L Hansen, "Unsupervised audio stream segmentation and clustering via the Bayesian information criterion," in *ICSLP*, 2000.

[6] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[7] Ananth N Iyer, Uchechukwu O. Ofoegbu, Robert E. Yantorno, and Stanley Wenndt, "Speaker modeling in conversational speech with application to speaker count," in *Proc. ICSLP*, 2006.

[8] S E Tranter and D A Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. On Audio Speech and Language Processing*, vol. 14, pp. 1557–1565, September 2006.

[9] Imran Ahmed and Sunil Kumar Kopparapu, "Speaker change detection in telephone speech," in *International Conference on Signals, Systems and Automation*, Vallabh Vidyanagar, India, December 2009.