A NOVEL METHOD FOR TWO-SPEAKER SEGMENTATION

Rashmi Gangadharaiah, B.Narayanaswamy§, N.Balakrishnan**

*Supercomputer Education Research Center (SERC), Indian Institute of Science, Bangalore
§ ISRI and ECE, Carnegie Mellon University, PA 15213
rashmi@mmsl.serc.iisc.ernet.in, muralib@cs.cmu.edu, balki@serc.iisc.ernet.in

Abstract

This paper addresses the problem of speaker based audio data segmentation. A novel method that has the advantages of both model and metric based techniques is proposed which creates a model for each speaker from the available data on the fly. This can be viewed as building a Hidden Markov Model (HMM) for the data with speakers abstracted as the hidden states. Each speaker/state is modeled with a Gaussian Mixture Model (GMM). To prevent a large number of spurious change points being detected, the use of the Generalized Likelihood Ratio (GLR) metric for grouping feature vectors is proposed. A clustering technique is described, through which a good initialization of each GMM is achieved, such that each state corresponds to a single speaker and not noise, silence or word classes, something that may happen in conventional unlabelled clustering. Finally, a refinement method, along the lines of Viterbi Training of HMMs is presented. The proposed method does not require prior knowledge of any speaker characteristics. It also does not require any tuning of threshold parameters, so it can be used with confidence over new data sets. The method assumes that the number of speakers is known apriori to be two. The method results in a decrease in the error rate by 84.75% on the files reported in the baseline system. It performs just as well even when the speaker segments are as short as 1s each, which is a large improvement over some previous methods, which require larger segments for accurate detection of speaker change points.

1. Introduction

In speaker identification applications, it is often assumed that the speech file contains data of a single speaker. However, in many applications such as identifying participants in a telephone conversation or in a conference, speech from different speakers is intermixed. Under such circumstances, classification of speech according to who is speaking becomes important. To do this, the beginning and end points of each speaker's voice are required. This paper provides a method that is novel and assumes no prior knowledge of the characteristics of the speakers.

Current methods to detect speaker changes are based on decoder based splitting, model based splitting or metric based splitting. Several cluster distances have been tested in [1,2]. Model based splitting uses models for each speaker, which are trained beforehand and is preferred when prior audio information is available about the speakers. Metric based splitting finds speaker changes based on maxima of some distance measure between two adjacent windows shifted along the speech signal. These segmentation algorithms suffer from a lack of stability since they rely on thresholding of distance values.

The GLR is the most computationally expensive distance measure but produces the best results [3], showing high and narrow peaks at speaker change points. The segmentation algorithm based on Bayesian Information Criterion (BIC), cannot detect two speaker changes closer to one another in time, as BIC has been shown to require longer speech segments [4,5]. The content based indexing proposed in [6,7] combines the GLR distance measure to detect the speaker change points and the BIC technique to refine the results in order to fight over-segmentation.

A good segmentation algorithm should meet the following requirements:

- It should not be overly sensitive to parameters such as window length, window overlap, and window shift so that they can be selected as a trade off between speed and accuracy.
- It should have the ability to detect the speaker change points accurately.
- It should result in segments with a single speaker.
- It should have optional refinement stages, which allow increased accuracy at the cost of speed.

In this paper, the window length has been taken as 1s, shifted by 0.5s, but the method was found to work satisfactorily with different values of these parameters. To meet the second and third condition, the GLR distance measure has been improved by further processing. An efficient refinement stage has been performed to solve the fourth requirement.

The rest of the paper is organized as follows. Section 2 describes the proposed method. Section2.1 describes the adaptive silence removal technique used. Section 2.2 describes the initial segmentation based on GLR metric, and Section 2.3 shows the initialization of GMMs for the speaker models, and Section 2.4 explains the refinement or Viterbi training of the clusters and models. Section 3 describes the evaluation set up, and compares the results obtained with the baseline system. Section 4 concludes and proposes some improvements to achieve better segmentation more efficiently.

2. Method

The proposed method consists of two major parts, the initial speaker change detection and the refining of models and change points. In many speech and speaker recognition tasks, models can be trained from a flat start, but in such cases, the models used for segmentation may converge to speech classes (say consonants and vowels) rather than to different speakers. To force the models to converge to models of the speakers, we need spectral features across longer segments, (at least 1-2s in length), to capture the long term speaker information but average out the short term speech information, that is, an initial segmentation that is more likely to contain data of a single speaker. An effective solution to this problem, using the GLR metric is proposed in section 2.3. An algorithm to derive good initial speaker models is described in section 2.4.

2.1. Silence Removal

A drawback of training HMMs on unlabelled data with silences is that some of the models may converge to these silence regions. The method used for silence removal in this paper is similar to the second method in the NIST stnr routine[8], a technique suggested by Ned Neuberg, Jordan Cohen and others. To save computation, only the first 5-10s is used for detecting the speech-silence threshold. A signal energy histogram is generated by computing the root mean squared (RMS) power, in decibels, over a 20ms window and then updating the appropriate histogram bin. The window is then shifted by 10 ms and the next power is computed. A plot of these power coefficients is seen to be bi-modal, with a sharp low energy silence mode and a flatter higher energy speech mode. The point of inflection between these two modes is the boundary between speech and silence. However, during the course of experimentation it was found that the threshold obtained at this point was too high and resulted in an increased number of deletion errors. The threshold was chosen to be the peak of the mode corresponding to the noise or silence region, since the method is robust to small silence regions.

2.2. Initial Segmentation

In this step, the dissimilarity between two contiguous windows of the parameterized signal is calculated. The GLR distance is computed for a pair of adjacent windows of the same size, and the windows are then shifted by a fixed step along the whole parameterized speech signal [7] as shown in Figure1. A large distance indicates change in speaker, whereas low values signify that the two portions of the signal correspond to the same speaker.

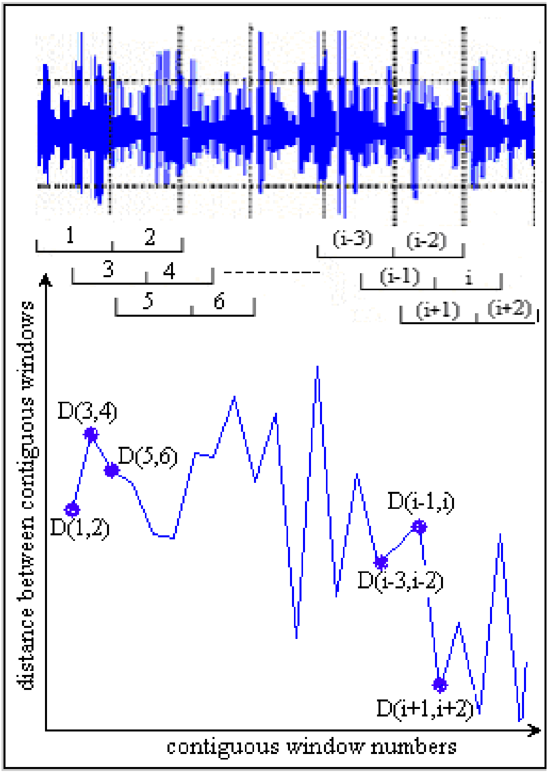

Figure 1: Distance Computation

Let D(i-1,i) denote the GLR distance between the (i-1)th and ith speech sub-segment. The initial segmentation is performed at peaks of this distance measure, which are above a threshold Th_i. The difficulty lies in setting the threshold, Th_i without any prior knowledge. A robust threshold can be set based on the previous N successive distances as follows[9],

$$Th_i = \alpha \frac{1}{N} \sum_{n=1}^{N} D(i - 2n - 1, i - 2n) \text{..........(1)}$$

Where, N is the number of previous distances used for predicting the threshold, and α is a coefficient used as an amplifier and it is set to 1.0. The threshold determined in this way was found to work satisfactorily. Thus the threshold is set automatically and need not be derived experimentally on each new data set.

The above procedure splits the audio data into a series of segments, SEG_1, SEG_2,……, SEG_n defined by the change points located at the peaks of the GLR distance metric.

2.3 Initialization of GMMs for the two speakers/states

2.3.1 Initial data for the first speaker:

The segment between the starting point of the conversation and the first detected change point is assumed to represent the first speaker's data. The feature vectors of this segment are modeled as a GMM, λ_a.

2.3.2 Initial data for the second speaker:

The feature vectors of each segment are modeled using one GMM per segment as $\lambda_2, \ldots \lambda_n$. The model, $\hat{S}$ having the minimum a posteriori probability for SEG_1, found using *(5),* is assumed to represent the second speaker's data.

$$\hat{S} = \arg\min_{2 \le k \le n} p(\lambda_k \mid SEG_1) \text{.........(2)}$$

By Baye's rule this becomes,

$$\hat{S} = \arg\min_{2 \le k \le n} \frac{p(SEG_1 \mid \lambda_k) p(\lambda_k)}{p(SEG_1)} \text{..........(3)}$$

where,

$p(\lambda_k)=1/n$, is probability of choosing a particular model, $p(SEG_1)$ remains same for models λ_k, and, the set of feature vectors $SEG_1 = \{s\vec{e}g_{11}, s\vec{e}g_{12}, \ldots s\vec{e}g_{1T}\}$

therefore,

$$\hat{S} = \arg\min_{2 \le k \le n} p(SEG_1 \mid \lambda_k) \text{......... .(4)}$$

$$\hat{S} = \arg\min_{2 \le k \le n} \sum_{t=1}^{T} \log\ p(s\vec{e}g_{1t} \mid \lambda_k) \text{.......... (5)}$$

This step results in the initialization of the parameters of GMMs representing the two states, to the values λ_a and the λ_k corresponding to the segment, $\hat{S}$. Thus each state is initialized to a segment which is most likely to be from a different speaker.

2.4 Refinement of the clusters and change points

The two clusters created in the initialization step are now used as the two *reference models (*λ_a and λ_b*)* and *(6)* is computed for all the segments created by the speaker change detection step. The objective here is to find the model which has the maximum aposteriori probability for each segment SEG_y, where, y=2,3..n. If the *reference model*, λ_a, shows a higher aposteriori probability compared to λ_b for SEG_y, then SEG_y will be labeled as the first speakers data otherwise then SEG_y will be labeled as the second speakers data. The segments are clustered based on the labels.

$$\hat{S} = \arg\max_{m=a,b} p(\lambda_m \mid SEG_y) \text{.........(6)}$$

$p(\lambda_m)=1/2$ (any segment is equally likely to belong to either speaker). This is similar to the Viterbi training algorithm for HMMs with all transition probabilities fixed and equal. A second iteration is performed to obtain clusters of high purity for the two speakers. This procedure can be repeated till convergence. The performance was found not to increase significantly beyond the fourth iteration.

3. Evaluation

A good segmentation task should detect the speaker change points accurately. There are two types of errors related to speaker change detection, an *insertion error* occurs when a speaker change is detected although it does not exist, a *deletion error* occurs when a true speaker change is not detected.

To compare the proposed approach with [6,7] the same evaluation tasks, on TIMIT and SWITCHBOARD (published by NIST and distributed by the LDC in 1992-3), are used. A set of 40 speakers (T) is taken from the dialect regions DR1 and DR2 from TIMIT. A conversation is obtained by concatenating sentences of 2s on average from two speakers taken from the set, T (clean speech). Two files from SWITCHBOARD, sw2005 and sw2007 are used in a second part of the comparison. Finally results are presented for some other SWITCHBOARD files to show that the proposed method is robust to noise/ silence regions and other variations in the speech signal.

12th order Mel-cepstral coefficients are computed for every frame of 16 ms length with 6.25ms shift for parameter extraction. For speaker change detection, the length of each window is set to 1s and shifted by 0.5s. A four component GMM with diagonal covariance is used to compute the GLR distance between two consecutive windows. For modeling the segments created, eight component GMM with diagonal covariance is used.

The method suggested in [6] involves two passes. In the first pass, a distance based segmentation is done using a 2s window shifted by 0.1s. The BIC is then used during the second pass to refine the previously detected change points. The number of deletion errors increases after the second pass, because when BIC is used for segmentation long speech segments are required. Table1 gives results obtained using the segmentation technique given in [6]. The first two and the next two columns correspond to the results obtained after the first and second pass respectively. The last two columns show the results obtained using the method proposed.

The resolution of the proposed method is Δt=0.5s, and therefore, if any segment boundary is hypothesized within the time interval, $t_0-\Delta t<t<t_0+\Delta t$ of the reference boundary t_0, it is regarded as correct. Utterances of less than 0.5s duration were not taken into account while marking the correct change points.

As can be seen from Table 1, the proposed method decreases number of errors on SWITCHBOARD by 84.75 %. Also, Table 2 shows that it performs as well in different conversations, in noise, with silence, or when the conversation is between speakers of same/different sex.

Table 1: Comparison of [6] and the proposed method on TIMIT and SWITCHBOARD.
I - number of Insertions, **D** - number of Deletions
F-female, **M**-male

Files	GLR and BIC as in [6]				proposed method	
	1st pass		2nd pass		I	D
	I	D	I	D		
TIMIT 29change points	26	3	9	7	**2**	**2**
TIMIT 27change points	23	3	9	7	**3**	**2**
sw2005 (M-M) 19change points	41	6	17	7	**0**	**2**
sw2008 (F-F) 30change points	31	17	18	17	**4**	**3**

Table2: Results obtained using the proposed method on SWITCHBOARD
I - number of Insertions, **D** - number of Deletions
F – female, **M** - male

Files	I	D	Characteristics
sw2001 (F-F) 34 change points	**3**	**3**	some background noise
sw2006 (F-M) 25 change points	**2**	**2**	large silences, noisy
sw2010 (F-M) 45 change points	**2**	**4**	some background noise
sw2015 (F-M) 24 change points	**0**	**3**	large silences, some background noise
sw2017 (F-M) 37 change points	**6**	**1**	large silences, some background noise
sw2018 (F-F) 65 change points	**8**	**2**	overlapped speech, background noise
sw2102 (F-M) 52 change points	**2**	**2**	noisy, few silences
sw2105 (F-M) 52 change points	**3**	**6**	noisy, large silences

4. Conclusion and Future work

In this paper, a novel speaker-based segmentation technique was proposed and tested on the TIMIT and SWITCHBOARD databases. The segmentation algorithm gave good results with few insertion or deletion errors. This method was applied for phone conversations.

For future work, we plan to modify this method to overcome problems like noise, overlapped speech, and also increase the resolution of the method. We also plan to extend this method to N-speakers, by extending the view of the data as a HMM with speakers as the hidden states.

5. Acknowledgements

We would like to thank Prof. Raj Reddy and Prof Richard Stern for their advice and guidance in all phases of this work. We would also like to thank Kishore Prahallad for his comments and suggestions that improved the quality of the paper considerably.

6. References

[1] Laurent Couvreur and Jean-Marc Boite, "Speaker Tracking in Broadcast Audio Material in the frame work of the THISL project", *Proc. of European Speech Communication Association ETRW Workshop on Accessing Information in Spoken Audio, Cambridge (UK), pp. 84-89, 1999.*

[2] Soonil Kwon, Shrikanth Narayanan, "Speaker Change Detection using a New Weighted Distance Measure", *IEEE International Conference on Spoken Language Processing, Denver, USA, vol.4, pp. 2537-2540, 2002.*

[3] H. Gish, M. Siu, R. Rohlicek, "Segregation of speakers for speech Recognition and Speaker Identification," *IEEE International Conference on Acoustics, Speech, and Signal Processing, Toronto, Canada, pp. 873-876, 1991.*

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

[5] Alain Tritschler and Ramesh Gopinath, "Improved speaker segmentation and segments clustering using Bayesian Information criterion", *Sixth European Conference on Speech Communication and Technology, Budapest, Hungary, pp. 679-682, 1999.*

[6] Perrine Delacourt, David Kryze, and Christian J. Wellekens, "Speaker-based segmentation for audio data indexing", *ISCA International Speech Communication Association, Cambridge, UK, pp.78-83, 1999.*

[7] Perrine Delacourt and Christian J.Wellekens, "Audio Data Indexing: use of Second Order Statistics for Speaker-Based Segmentation", *International Conference on Multimedia Computing and Systems, Florence, Italy, vol.2, pp. 959-963, 1999.*

[8]Casimir Wierzynski and Jon Fiscus, "stnr.doc" *included with the NIST SPeech Quality Assurance (SPQA) Package Version 2.3 AND Speech File Manipulation Software (SPHERE) Package Version 2.5.*

[9] Lie Lu, Hong-Jiang Zhang, and Hao Jiang, "Content Analysis for Audio Classification and segmentation*", IEEE transactions on speech and audio processing, Vol. 10, No. 7 pp.504-516, 2002.*