Identifying Speaker Change Points in Real Telephone Speech

Imran Ahmed, Sunil Kopparapu
TCS Innovation Lab
Tata Consultancy Services Ltd.
ahmed.imran@tcs.com, sunilkumar.kopparapu@tcs.com

***Abstract*—Speaker change point detection is a crucial step in voice analytics in general and in speaker segmentation process in particular. Speaker change detection in telephone conversations has gained importance in recent times for intelligence agencies. While techniques exist for speaker change detection in literature for audio broadcast and meeting, there is not much work done for telephone conversations because telephone speech has several challenges like much faster speaker change rate, variation of speaking style by a talker, and presence of non-speech sounds, crosstalk and babble. In this paper we propose a method to detect change point in real telephone conversations. Experiments on real world telephone conversations show the effectiveness of the proposed system.**

***Index Terms*—speaker change detection; telephone conversation segmentation.**

I. Introduction

Automatic identification of the number of speakers and speaker change detection in telephone conversations is a very important problem in voice analytics in general and for speaker identification and verification in particular. A number of methods for automatic segmentation and diarization of audio data in general and speaker segmentation in particular have been proposed in literature. Almost all the systems have two major components namely (a) acoustic change detection (ACD), also called speaker change detection and (b) speaker clustering (SC). Speech detection and gender segmentation are some of the other optional components depending on the objective of the systems. ACD component looks at acoustic properties to identify speaker change points which might also include speech non-speech change points. The most common approach for change detection is to first divide the full conversation audio into large number of small overlapping window frames and then look at adjacent frames and calculate dissimilarity metric between the two adjacent frames. A threshold decides whether the frames originate from the same or a different source. Depending on the ultimate goal, this is usually followed by the speaker clustering component where homogeneous frames not necessarily contiguous in time are grouped together using agglomerative hierarchical clustering [1], [2] or using segment adapted speaker/background model based approaches [3]. The output of clustering stage may be again used to refine the initial change detection performed by ACD.

While this works for broadcast audio or audio data from meetings, the performance is hampered for live telephone conversations in terms of purity of the final segments. This is because (a) telephone conversations have a much faster speaker changing rate (and hence many true small length segments), (b) variation of speaking style by a talker, and (c) presence of non-speech sounds and crosstalk (which may be smaller than 1s). As a result the change point detection has several errors; both false alarms and missed recognitions. As reported in [4] performance of speaker segmentation systems is affected by the speaker change rate and speaker turn durations. This work studies the characteristics of speaker diarization audio data, particularly characteristics which are associated with higher DER (Detection Error Rate) [4].

There have been efforts in literature to apply traditional speaker change detection and segmentation techniques to telephone conversations. The approach in [5] is capable of detecting longer segments (> 4 secs) with miss rates of the order of 10% and confusion rates 2% or less but does not resolve short (< 2 secs) or overlapping segments very well. [6] is based on approach in [5], with an assumption that the telephone conversation involves two speakers. The authors in [6] provide a stopping criterion which helps improve the purity at the clustering during segment refinement stage. Another approach in [7] uses small overlapping windows (1sec) for unsupervised segmentation of two speaker conversation. However, it is based on the assumption that one of the speakers initiates the conversation and speaks continuously for at least 1 sec and that the 1 sec segment farthest to first speaker is the second speaker. In reality the 1 sec assumption for first speaker does not hold true for many telephone conversations and the second speaker assumption does not hold in case of noisy telephone conversations. And also clustered segments may not be able to refine change points beyond the accuracy of the change point detection used.

In this paper we discuss an unsupervised change point detection in telephone conversations which is able to reliably detect changes smaller than 1 sec. The proposed system uses smaller length non overlapping windows (1 sec or less) for change detection in telephone conversations. Small window size ensures that none of the actual change points are not missed. The rest of the paper is organized as follows: Section II presents the proposed change detection method suitable for telephone speech conversations; Section III describes the experimental setup and results. Section IV gives the conclusion.

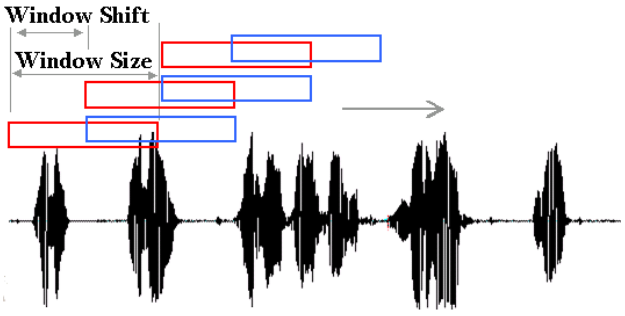

Fig. 1. Adjacent window comparison using overlapping windows.

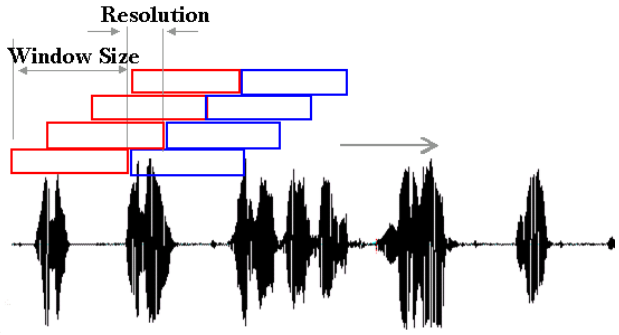

Fig. 2. Adjacent window comparison using non overlapping windows.

II. Change point detection

Change point detection is the process of identifying changes in some characteristics of the speech. In a telephone conversation these change in characteristics can be assumed to be that of detection of change from silence to speech, speech to silence, one speaker to another. Typically in telephone conversations, in addition to the noise and cross talk which is all too common, there are several instances when more than one person is speaking at the same time. Additionally, the conversation can be skewed in the sense that the time duration of speech of one person could be very small compared to that of the other person. These are some of the challenges that one faces in telephone conversation compared to audio broadcast. In this section we describe the general and common processes along with our contribution adopted for automatic change point detection in telephone conversation.

A baseline change point detection technique used in speaker segmentation systems [1] is to divide the entire speech into small overlapping windows of size 25 secs. Adjacent windows are then compared[1] to 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[2]) are compared to determine if these windows of speech have similar characteristics or 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 (see Figure 1). Typical telephone conversations have a large number of short speech segments ($<$ 1, 2 sec). Thus, for a typical window size in the range 2 to 5 sec the number of missed change point is high, even with a small window shifts. On the other hand for small window size and small window shift, a large number of non-change points are detected as change points (false alarms). In order to enable the detection of smaller speech segments, independent of the window shift, we propose an alternate windowing technique using small non overlapping windows as shown in Figure 2. In the proposed approach, small non overlapping adjacent windows (red and blue boxes in Figure 2) are compared to enable the detection of smaller speaker segments in telephone conversation. The window size is kept small (less than 1 sec) to detect short speaker segments. Compared to the overlapping window method the window shift is equal to window size (non overlapping windows) and the change point is marked at the end of first window. The adjacent windows together, slide by an amount called resolution (see Figure 2) along the time axis. It is referred as the resolution because it determines the localization errors in the detected change point.

The following sub-sections describe the proposed change point detection system in detail.

A. Front-end Processing

The telephone conversations are 8 kHz sampled and are directly used for speech feature extraction. MFCC (Mel Frequency Cepstral Coefficient) and LSF (Line Spectral Frequency) are commonly used speech features for speaker segmentation [1]. It has been documented in literature that LSF speech parameters perform better for speaker segmentation [7]–[9], and hence for the comparison purpose we use only LSF features. The LSF features are calculated over 30 ms frames of the audio stream.

B. Detect potential speaker change point

For each window, the extracted LSF features are modeled as a single Gaussian distribution with a mean μ and covariance Σ. KL (KullbackLeibler) divergence measure [8] is used to measure the dissimilarity of the Gaussian distributions i.e. the speech in one window and speech in adjacent window. KL distance between i-th and j-th window is defined as

$$\begin{aligned} D(i,j) &= \frac{1}{2}(\mu_j - \mu_i)^T(\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_j - \mu_i) \\ &+ \frac{1}{2}tr\left((\Sigma_i^{1/2}\Sigma_j^{-1/2})(\Sigma_i^{1/2}\Sigma_j^{-1/2})^T\right) \\ &+ \frac{1}{2}tr\left((\Sigma_i^{-1/2}\Sigma_j^{1/2})(\Sigma_i^{-1/2}\Sigma_j^{1/2})^T\right) \end{aligned} \quad (1)$$

[1]Each window is preprocessed to extract features(MFCC,LSF or other depending on system implementation.

[2]In color print.

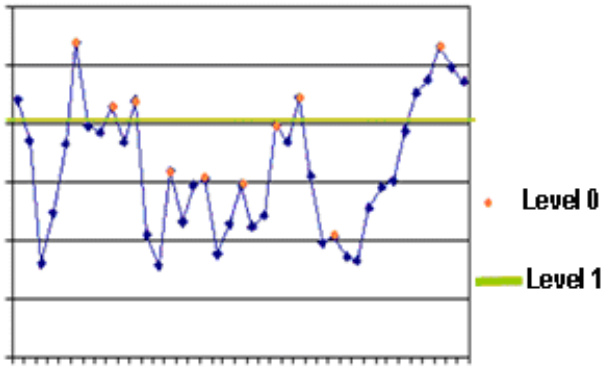

Fig. 3. Level 0 and Level 1 filtering of initial change points.

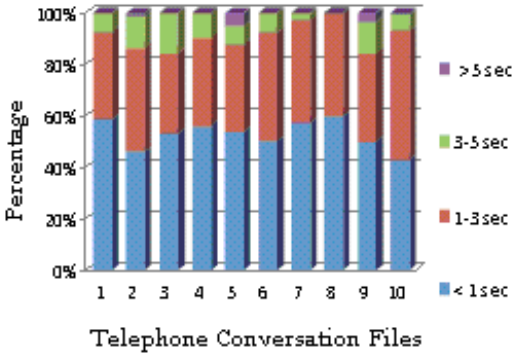

Fig. 4. Distribution of speaker segment lengths for 10 telephone conversations.

where tr is trace of the matrix.

C. Change Point Selection

All the identified change points are filtered to determine the correctness of the detected change point using the following techniques.

1) Silence regions: Telephone conversation has zones of silence regions which typically mark a change in the speaker. An energy based approach is used to extract regions of silence in telephone speech. We process the entire telephone conversation and eliminated any change point detected in the silence regions as being false alarms. Additionally, it was observed that in real telephone conversations, silence actually indicate a potential speaker change. Hence, we also marked beginning and end of silence as valid change points.

2) Filtering based on a threshold: In this step a potential speaker change is selected between the ith frame and the adjacent (i+1)th frame, if the following conditions are satisfied:

$$\begin{aligned} \text{Condition 1} &: D(i,i+1) > D(i+1,i+2) \\ \text{Condition 2} &: D(i,i+1) > D(i-1,i) \\ \text{Condition 3} &: D(i,i+1) > T \end{aligned} \tag{2}$$

where T is a threshold calculated at run time and D(i,j) is the metric that measures the similarity score between the ith and jth frames. Condition 1 and Condition 2 guarantee that a local peak exists, and the Condition 3 can prevent very noisy low peaks from being detected. This condition though simple helps in determining change points with reasonable accuracies.

D. Experimental Analysis

To test and analyze the performance of the proposed change point detection system, we choose 10 real telephone conversations ranging from 1 to 4.5 minutes duration. Note that these telephone conversations were real in the sense that the people on the call were unaware of the recording. The recordings were captured from different telephone exchanges, giving the telephone conversations the required variability in terms of not being biased by a particular type of recording environment. Additionally, as is the case in most of the telephone conversation, there were only two speakers in the conversation but speaking for varying length of time. The distribution of speaker segment lengths for each of these conversations is as shown in Figure 4. As observed in the distribution graphs, telephone conversations have a majority of the segments of duration less than 1 sec and 3 sec.

We choose different window sizes to evaluate the proposed change point detection. We also considered both overlapping and non overlapping windows as discussed in Section II. The analysis window size was kept in the range of 0.75 to 4 s and the window was shifted by 0.2 s and more (this enables detection of small segments). In all our experiments an energy based silence detection algorithm was used to detect silences greater than 0.5 s. These silence regions were used for removing falsely detected change points as discussed in Section II.

In order to determine the accuracy of the detected changes the points detected by the system are compared to the manually marked change points (ground truth).The performance of change detection system can be measured using two figures of merit. One is false alarm rate (FAR) and miss detection rate (MDR) namely,

$$FAR = FA/(FA+GT) \tag{3}$$

$$MDR = MD/GT \tag{4}$$

where FA is the count of change points detected by the change point detection process but are not true change points (false alarm), while misses (MD) are those change points that exist in the ground truth but have been missed detection by the change detection process. GT is the ground truth i.e. the actual number of change points present.

The other measure (popularly used in information retrieval literature) is precision (PRC) and recall (RCL) rates defined as:

$$PRC = CFC/DET \tag{5}$$

$$RCL = CFC/GT \tag{6}$$

where DET is the total number of change points detected by the system and CFC = DET $-$ FA is the number of correctly

TABLE I
AVERAGE FALSE ACCEPTANCE (FAR) AND MISS DETECTION (MDR) RATES

Window length	NonOverlapping window		Overlapping window	
	FAR	MDR	FAR	MDR
0.75 s	0.35	**0.08**	0.38	0.09
1.0 s	0.35	0.10	0.22	0.20
2.0 s	0.33	0.13	0.14	0.29
4.0 s	0.29	0.14	0.07	0.34

TABLE II
AVERAGE PRECISION (PRC) AND RECALL (RCL) RATES

Window length	NonOverlapping window		Overlapping window	
	PRC	RCL	PRC	RCL
0.75 s	0.66	0.91	0.64	0.91
1.0 s	0.68	0.90	0.81	0.80
2.0 s	0.69	0.87	0.90	0.71
4.0 s	0.73	0.86	0.99	0.66

found change points. In all our experiments we identify a miss in change point if:

$$(detectedchangept. - actualchangept.) <= tolerance$$

Note that in case of non overlapping window, the tolerance is the resolution while in the case of overlapping window the tolerance is the window shift itself. Tolerance helps in not only checking the accuracy of change point detection but also gives an indication of purity of the segments. It is observed that the missed change points are of two types based on their importance, namely, missing a speaker change point is more expensive than missing a non speaker change point or a speaker change point with short speaker duration (< 0.75 s).

TABLE 1 shows the average FAR and MDR for the 10 telephone conversations tested using both overlapping and non overlapping windows discussed in Section II A, for different window lengths. Similarly, TABLE 2 shows the precision and recall rates.

It can be seen that the MDR for non small overlapping windows is small and around 50% of that using overlapping windows. The MDR increases as the window size is increased. FAR reduces with the increase in window size and also the overlapping window method has a comparatively low FAR. The low FAR in both the cases is mainly because the DET (total number of points detected by algorithm) itself is low. This can be observed in precision and recall rates of overlapping windows of 2 and 4 sec lengths.

III. CONCLUSION

In this paper we addressed the problem of change point detection to enable speaker segmentation in telephone conversations. We proposed the use of smaller windows sizes for change point detection and showed that the use of smaller length non overlapping windows increases the precise of detection of change points, which translates to pure and accurate change point detection in telephone conversations. The smaller length non overlapping window method outperforms the overlapping window method for detection of very small segments in telephone conversation.

REFERENCES

[1] T. S.E. and R. D.A, "An overview of automatic speaker diarization systems," *IEEE Trans. On Audio, Speech and Language Processing*, vol. 14, p. 15571565, September 2006.

[2] S. S. Chen and P. S. Gopalakrishnam, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *In Proc. DARPA Broadcast News Transcription and Understanding Workshop , Lansdowne*, 1998, p. 127132.

[3] G. Rashmi and N. Balakrishnan, "A novel method for two-speaker segmentation," in *In INTERSPEECH-2004*, 2004, pp. 2337–2340.

[4] M. N. and W. C., "Nuts and flakes: a study of data characteristics in speaker diarization," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2006 Proceedings.*, 2006, pp. 1017–1020.

[5] R. A. E., G. Allen, L. Zhu, and P. S., "Unsupervised speaker segmentation of telephone conversations," in *ICSLP-2002*, 2002, pp. 565–568.

[6] Z. Xin, C. M. A., and L. Sung, "Acoustic change detection and segment clustering of two-way telephone conversations," in *EUROSPEECH-2003*, 2003, pp. 2925–2928.

[7] Adami, A. K. S.S., and H. H., "A new speaker change detection method for two-speaker segmentation," in *Proceedings. IEEE International Conference on Acoustics, Speech, and Signal Processing -ICASSP 02*, 2002, pp. IV–3908 – IV–3911.

[8] P. Delacourt and C. Wellekens, "Distbic: A speaker-based segmentation for audio data indexing," *Speech Communication*, vol. 32, p. 111126, September 2000.

[9] L. Lu and H.-J. Zhang, "Real-time unsupervised speaker change detection," in *Proceedings to 16th International Conference on Pattern Recognition*, 2002, pp. 358–361.