Speaker Change Detection using Excitation Source and Vocal Tract System Information

Mousmita Sarma † ‡, Sree Nilendra Gadre *, Biswajit Dev Sarma * and S. R. Mahadeva Prasanna * ‡

*Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati, Guwahati-781039 , Assam, India

† Department of Electronics and Communication Engineering
Gauhati University, Guwahati-781014, Assam, India

‡ Speechwarnet (I) Pvt. Ltd,
TIC, IIT Guwahati, Guwahati-781039, Assam, India
Email: mousmita@speechwarenet.com, {prasanna, sree, s.biswajit}@iitg.ernet.in

***Abstract*—The speaker change information in speech is due to both vocal tract and excitation source information. In this work, the excitation source information is extracted by computing cepstral features from the zero frequency filtered speech (ZFFS) signal. The vocal tract system information is extracted by computing cepstral features from the speech signal. The speaker change evidences obtained from these two feature sets are combined and observed that they contain complementary information for speaker change detection. The popular distance metric based algorithms, Bayesian Information Criteria (BIC) and Kullback Leibler Divergence (KLD) are used to detect the speaker change evidences. The Miss Detection Rate (MDR) of BIC based algorithm using cepstral features obtained from speech is 24.18 % and from ZFFS is 25.92%, respectively. When the two sets of evidences are combined, the MDR reduces to 15.89%. Similarly, individual MDR of KLD based algorithm from speech and ZFFS are 32.24% and 45.17%, respectively, where as the combination reduces the MDR to 19.67%. Experiments are also performed with noisy speech signal and similar reduction of MDR is observed. This demonstrates the usefulness of cepstral features from the excitation source signal for reducing MDR.**

I. Introduction

Speaker segmentation is the first phase of a speaker diarization system, which is followed by a speaker clustering phase. Speaker segmentation aims at finding the speaker change points in an audio stream. Therefore, speaker change detection is the crucial part in the development process of a speaker diarization system [1] [2]. Various methods are proposed in the literature for this purpose like distance metric based, model based, silence detection based etc. along with some hybrid methods. Distance metric based methods like Bayesian Information Criteria (BIC) [3] [4] [5], Kullback-Leibler Divergence (KLD) [5] [6] [7] and Gaussian Divergence or symmetric Kullback-Leibler-2 Divergence (KLD2) [8] [9] are the most popular and widely used. Another popular approach of speaker segmentation is the model based segmentation, where a set of models are derived and trained for different speaker classes from a training database. Maximum likelihood selection is then used to classify the audio stream based on these models and boundaries between the models are decided as change points [2] [10]. Primary difference between model based approach and distance metric based approach lies in the requirement of prior knowledge in the first one to initialize the speaker models. Silence detection based methods are also used for speaker segmentation where it is assumed that a silence region is present between two speakers speech. But other techniques are always required to validate the change point detected by silence detection based methods. This is because, silence may not always present between the utterances of two speakers [1]. Further hybrid methods combining all these approaches are also proposed in the literature [5] [11] [12].

978-1-4799-6619-6/15/$31.00

In all the above methods, for proper detection of speaker change points the most crucial component is the features extracted from the acoustic signal. The feature set should convey information specific to the speakers in the conversation so that they can provide appropriate discrimination for speaker modeling. Further features should be robust against noise and distortion. Mel frequency cepstral coefficients (MFCC) with their first and/or second derivatives [3] [13] [14], short-time energy [15], zero-crossing rate (ZCR) [16] and perceptual linear prediction (PLP) cepstral coefficients [17] are the most common features used for speaker diarization. Out of these, MFCC is the most dominant method.

Speech signal carries both linguistic as well as speaker specific information along with the message to be conveyed. Speech signal production is described as a filtering process in which a speech sound source excites the vocal tract filter. Thus both vocal tract system and excitation source bears speaker specific information. The conventional MFCC features provides speaker-specific vocal tract information. However none of these commonly used feature extraction methods provides features specific to excitation source information. This work explores the usefulness of speaker specific information obtained from the excitation source signal so that the miss detection rate (MDR) can be reduced in speaker change detection. The excitation source signal is obtained by filtering the speech signal using zero frequency filter (ZFF) [19]. Here MFCC features are extracted from zero frequency filtered speech (ZFFS) signal . The ZFFS preserves excitation source information around zero frequency, thus avoids vocal tract information. Another advantage of using ZFFS is that it will not contain high frequency speech components mixed

with external noise. To obtain compact feature representation of ZFFS, MFCC of ZFFS are computed and used as features.

The usefulness of proposed feature is experimented separately in two distance metric based method of speaker change detection namely, BIC and KLD. Further in order to check noise robustness, white noise is added in the raw speech signals. The consistency of reduced MDR using the combination of both the features will be observed in both BIC and KLD for both clean and noisy speech signal. These studies will help in establishing the different nature of speaker change information present in both cases and also its robustness.

IITG-MV Phase III Speaker Recognition Database Part II (two speaker recognition) is used to evaluate the performance of the algorithm. It is a conversation mode database where every speaker spoke in conversational style over a conference call. It has the variabilities like multi-environment, multi-sensor and multi-lingual [18].

The rest of the paper is organized as follows: Section II provides a brief description of the distance metric based BIC and KLD method for speaker change detection. Computation of speaker specific excitation source based feature is described in Section III. Section IV includes the experimental details and results. Section V concludes the description.

II. Distant Metric based speaker change detection

In popular distance metric based speaker change detection, the distance measure gives the dissimilarity of feature vectors within the two windows. Here a distance metric between every two consecutive analysis segments is used as a decision measure to determine the change points.

A. Bayesian Information Criteria (BIC)

BIC is a model selection criteria derived from the generalised likelihood ratio (GLR). BIC searches for change points within a window using penalized likelihood ratio test of whether the feature vectors in the window is better modeled by a single distribution or two different distributions. Therefore at the event of speaker change, feature vectors within the window will be better modeled by two different distributions [4] [5].
Here two models are compared. One models the data as two multivariate Gaussians and the other models the data as just one Gaussian.
Consider $X = x_i \in R^d, i = 1, ..., N$ as the d-dimensional sequence of feature vectors and assume X is drawn from an independent multivariate Gaussian process:

$$x_i \sim N(\mu_i, \Sigma_i)$$

where, μ_i is the mean vector and Σ_i is the full covariance matrix.
The maximum likelihood ratio statistics is

$$R(i) = Nlog \mid \Sigma \mid -N_1 log \mid \Sigma_1 \mid -N_2 log \mid \Sigma_2 \mid$$

where, Σ, Σ_1 and Σ_2 are the sample covariance matrices from all the data, from $x_1, \cdots, x_i$ and from $x_{i+1}, \cdots, x_N$, respectively.
The difference between the BIC values of these two models can be expressed as

$$BIC(i) = R(i) - \lambda P$$

where λ is the penalty factor and

$$P = \frac{1}{2}(d + \frac{d(d+1)}{2})logN$$

B. Kullback-Leibler Divergence (KLD)

KLD is a distance measure for finding the differences among the two distributions. Feature vectors belonging to a particular speaker form a separate cluster. Therefore at the speaker change point, KLD is high compared to other points. The KLD between two multivariate Gaussian distributions $N(\mu_0, \Sigma_0)$ and $N(\mu_1, \Sigma_1)$ is [6] [5]

$$D_{KL(N_0 \| N_1)} = \frac{1}{2}(tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T \Sigma_1^{-1}(\mu_1 - \mu_0) - k - ln(\frac{det\Sigma_0}{det\Sigma_1}))$$

III. Computation of speaker specific excitation source based feature

In this work, a new feature set is proposed, which provides information about the excitation source of the speech signal. Basically mel frequency cepstral analysis is performed on the ZFFS. When the speech signal is passed through the ZFF, energy only around zero frequency is preserved at the output. Since the excitation to the vocal tract system is basically impulse like excitation [19], therefore the output of ZFF consists of excitation source information. The ZFFS is robust to high frequency noise, since only low frequency components are preserved. The ZFFS can be computed from the speech signal as follows [19]:

1) Difference the speech signal $s[n]$ to remove any DC component introduced by the recording device.
$$x[n] = s[n] - s[n-1]$$
2) Pass the differenced speech signal $x[n]$ through a cascade of two ideal zero-frequency resonators. That is
$$y_0[n] = -\sum_{k=1}^{4} a_k y_0[n-k] + x[n]$$
where , $a_1 = -4$, $a_2 = 6$, $a_3 = -4$ and $a_4 = 1$.
3) Compute the average pitch period using the autocorrelation of 30 ms speech segments
4) Remove the trend in $y_0[n]$ by subtracting the local mean computed at each sample. The resulting signal
$$y[n] = y_0[n] - \frac{1}{2N+1}\sum_{m=-N}^{N} y_0[n+m]$$
is the ZFFS signal. Here $2N + 1$ corresponds to the number of samples in the window used for mean subtraction.

After that mel frequency cepstral analysis is performed on the ZFFS signal. In the process of computing MFCC, speech signal is analyzed with short time fourier transform (STFT) and the spectrum is passed through the mel filter bank. Thus spectral values are grouped together in critical bands and weighted according to the triangular weighting function. Cepstrum analysis is performed on the mel frequency spectrum obtained this way.

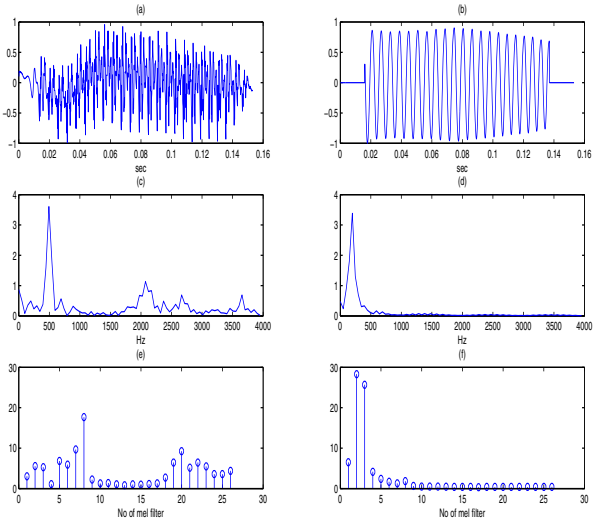

Fig. 1. (a) Voiced speech segment, (b) ZFFS of voiced speech segment in (a), (c) Magnitude spectrum in linear scale of voiced speech segment in (a), (d) Magnitude spectrum in linear scale of ZFFS of voiced speech segment in (a), (e) Energy after passing through mel filetr bank for voiced speech segment in (a), (f) Energy after passing through mel filetr bank for ZFFS of voiced speech segment in (a)

Fig. 1(a) and (b) represent a voiced speech segment and its ZFFS. Fig. 1(c) and (d) represent their respective magnitude spectra in linear sacle. As can be seen from the spectrum that energy only around very low frequency are preserved after filtering at zero frequency. Fig. 1(e) and (f) represents energy after passing through the mel filter bank which are also different for speech and ZFFS. In this work it has been observed that if the spectra of ZFFS signal is passed through the mel filter bank, the filter bank energy at the output is different from that of speech signal. Hence the mel frequency cepstrum analysis of ZFFS signal provides cepstral coefficients which can be considered as a feature set obtained using the excitation source information.

IV. Experimental Results and Discussion

BIC and KLD based speaker change detection algorithms are performed to check the effectiveness of the new feature set. Initially the speaker change detection is done separately for MFCC of speech and MFCC of ZFFS. Later the evidences obtained from both the feature sets are combined to form a hybrid set of evidences. The performance of speaker change detection algorithms are evaluated by computing the Miss Detection Rate (MDR) and False Alarm Rate (FAR) based on the manually marked change points provided with speaker database. MDR and FAR are defined by

$$MDR = \frac{\text{No. of miss detections}}{\text{No. of actual speaker boundaries}}$$

$$FAR = \frac{\text{No. of false alarms}}{\text{No. of detected speaker boundaries}}$$

FAR and MDR both should be lowest in order to achieve best performance [1].

A. Database used for the work

IITG-MV Phase III Speaker Recognition Database Part II (two speaker recognition) [18] is used here to evaluate the performance of the algorithm. This database is created in Indian scenario where there is a wide variability present in the languages, styles, environmental conditions and sensors. The Phase-III data is collected in telephone network. It contains speech data collected over conference call mode between two speakers when a facilitator connects the call in. The variabilities present in the Phase-III database are as follows [18]:

- Multi-environment: Speech data recorded by conversing in all kinds of practical environments possible like coffee shops, working places, rooms, laboratories etc.
- Multi-sensor: Speech data recorded over different mobile handsets at sampling frequency of 8 kHz.
- Multi-lingual: Every speaker spoke either in English language or his/her mother tongue. Languages present are Assamese, Hindi, English, Telegu, Kannada, Malayalam, Oriya, Bengali, Gujarati, Marathi, Bhojpuri, and Marwari.
- Conversation style: Every speaker spoke in conversational style over a conference call.

Ground truth speaker change marks are provided along with the database, marked with human effort by listening the speech signal using headphone and observing the waveform in wavesurfer. It is observed that the average speaker turn duration in the database is 4.5 sec. This statistics is computed from the manually marked speaker change points. Out of the 100 files of 10 minutes duration, there are a total of 12931 speaker change points in a average of 4.5 sec turn duration.

B. Peak Picking method

In distance metric computation a high distance value indicates a possible acoustic change whereas a low value indicates that two portions of signal corresponds to the same acoustic environment. The distance contour obtained from BIC and KLD consisted of various high and low peaks indicating both speaker and sound unit changes. Hence an efficient peak picking method is required to detect speaker change related peaks. Traditionally, in case of BIC if $\max_i BIC(i) > 0$, a local maximum of $BIC(i)$ is considered to be obtained and time i is considered to be a speaker change instant [4]. Similarly in case of KLD any local maxima greater than certain threshold is considered as a speaker change point. In our work a new peak picking method is considered which is same for both BIC and KLD. The steps involved in the peak picking method are as follows.

- Step 1: The distance contour obtained from BIC and KLD are smoothed by convolving with a hamming window of length 500 ms.
- Step 2: In the smoothed contour mean of the distance values is computed. Then an axis is set at the mean value and maximum distance value between a consecutive positive and negative mean crossing point is

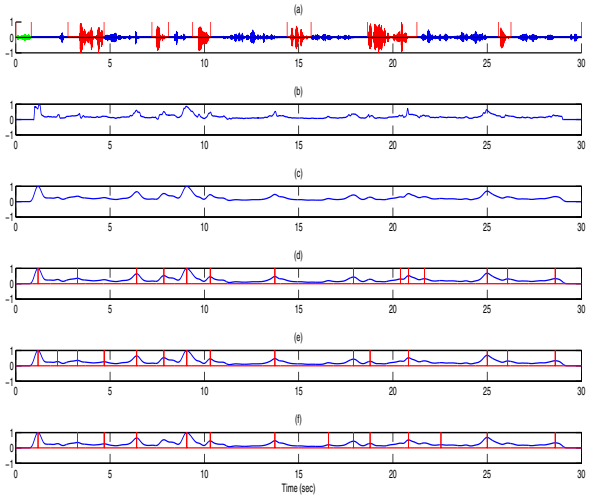

Fig. 2. (a) Speech signal with color change indicating ground truth marking, (b) Distance contour obtained from KLD, (c) Hamming window smoothed KLD distance contour, (d) Peak points obtained by considering mean+0.05 crossing points, (e) Peak points obtained by considering mean crossing points, (f) Peak points obtained by considering mean-0.05 crossing points

marked. That means peak points are located at the maxima between every pair of positive and negative mean crossing points. Thus a set of peak points is obtained. In the same way two more axis are set at $mean+0.1$ and $mean-0.1$ for BIC and $mean+0.05$ and $mean-0.05$ for KLD. Thus two more sets of peak points are derived.

- Step 3: In the next step all the three sets of peak points are combined together. Unique peak points from the combined sets are extracted removing any repetitions and hypothesized to be the speaker change evidences.

Fig. 2 shows the algorithm level plots for KLD. Here, Fig. 2(a) shows a 30 sec speech signal of two speakers along with ground truth marking indicated by red vertical lines. Fig. 2(b) shows the basic KLD contour and Hamming window smoothed KLD contour is shown in Fig. 2(c). Peak points obtained at three levels of peak picking are shown in Fig. 2 (d), (e) and (f) by red vertical lines. It can be seen here that at the three levels of peak picking, different peaks are obtained along with some repetition.

C. Speaker change detection using vocal tract information

The 13 MFCCs are extracted from the speech signal at every 20 ms frame with a shift of 10 ms. Number of filters in the mel filter bank is 26. In both BIC and KLD distance measurement is performed between the feature vectors coming inside two consecutive analysis windows, shifted along the MFCC feature space and distance contour is obtained. The analysis window size considered here is 1000 ms with a shift of 10 ms. The distance contour thus obtained is then resampled to the original speech signal length and normalized to the maximum value. Peaks related to speaker change points are next detected using the peak picking method described in Section IV-B. Performance is evaluated over the whole database separately for clean and noisy cases for both BIC and KLD. To obtain noisy speech signal 10 dB white noise is added. Performance of speaker change detection using MFCC features obtained from clean and noisy speech for BIC and KLD are given in Table I.

TABLE I. PERFORMANCE OF SPEAKER CHANGE DETECTION ALGORITHMS

Sl. No	Speech Signal	Algorithm	Feature	MDR (%)	FAR (%)
1	Clean	BIC	MFCC of Speech	24.18%	66.34 %
			MFCC of ZFFS	25.92%	69.02 %
			Combined	15.89%	72.96 %
		KLD	MFCC of Speech	32.24%	71.56 %
			MFCC of ZFFS	45.17%	69.59 %
			Combined	18.37%	71.77%
2	Noisy	BIC	MFCC of Speech	26.86%	67.35%
			MFCC of ZFFS	30.51%	69.01%
			Combined	19.67%	75.24 %
		KLD	MFCC of Speech	35.32%	70.96%
			MFCC of ZFFS	35.28%	70.17%
			Combined	19.99%	74.78%

D. Speaker change detection using excitation source information

13 MFCCs are extracted from the ZFFS signal at every 20 ms frame with a shift of 10 ms using the same mel filter bank. After that BIC and KLD based speaker change detection is performed with analysis window of 1000 ms with a shift of 10 ms. The distance contour thus obtained from the excitation source information are resampled and normalized and subsequent peak detection is performed. Performance of speaker change detection using MFCC features obtained from ZFFS signal for BIC and KLD are given in Table I. The combination of MFCC of speech and MFCC of ZFFS of speech provides additional advantage in case of noisy speech signal. The vocal tract information is normally mixed with noise. But the affect of noise in excitation source information is very negligible. As can be observed from Table I, in case of noisy speech individual performance of MFCC of ZFFS is better than the MFCC of speech.

E. Combination of speaker change evidences

The speaker change evidences obtained from both MFCC of speech and MFCC of ZFFS are combined together. Initially if any change point is found to be repeated, then that is removed and only unique change points are considered. However, some nearby change points are observed representing the same speaker change which creates additional false alarm. Therefore a merging operation is done. If two change points are separated by less than 250 ms, then they are merged to their center point. Thus the hypothesized set of speaker change evidences are obtained, which are detected by speaker specific features obtained from both excitation source and vocal tract system information.

F. Discussion

In this section we have discussed the various issues related to the experimental results of the current work. Fig. 3 shows the distance contour and hypothesized speaker change points obtained from BIC for MFCC of speech and MFCC of ZFFS. Fig. 3(a) shows the 30 sec of a raw speech signal obtained from

the database, where two speakers are conversing in Hindi and English. Here change of color indicates the change in speaker. Further exact location of ground truth marking is indicated by red vertical lines in Fig. 3(a). The hamming window smoothed distance contour is shown in Fig. 3(b) along with the hypothesized speaker change points indicated by green vertical lines. Fig. 3(c) shows the ZFFS obtained from the raw speech signal of Fig. 3(a). Fig. 3(d) shows the Hamming window smoothed distance contour obtained from MFCC of ZFFS, where the green vertical lines indicates the hypothesized speaker change points. The combined set of speaker change points obtained from both the features is shown in Fig. 3(e) . Fig. 4 shows the respective distance contours and hypothesized speaker change points obtained from BIC for MFCC of noisy speech and MFCC of ZFFS of noisy speech. From Fig. 4(b), it can be observed that the BIC distance contour obtained from MFCC of noisy speech shifted down due to the addition of noise in the signal. This is because the vocal tract information mixed with noise very easily. But as can be seen from Fig. 4(d), the BIC distance contour obtained from MFCC of ZFFS has no such shifting of position, since the source information is less affected by high frequency noise. But in our algorithm pick peaking is done with reference to the mean of the distances, which helps to overcome the shifting problem of the BIC contour for MFCC of speech.

Performance of speaker change detection for BIC and KLD using MFCC features obtained from clean speech and ZFFS signal obtained from clean speech and the combined method are given under the serial no.1 of Table I. Similarly performance of speaker change detection for BIC and KLD using MFCC features obtained from noisy speech and ZFFS signal obtained from noisy speech and the combined method are given under the serial no.2 of Table I. The MFCC of clean speech basically contains high frequency vocal tract information and MFCC of ZFFS obtained from clean speech basically contains low frequency source specific information. Therefore it is observed that in case of clean and noisy speech for both BIC and KLD, MFCC of speech and MFCC of ZFFS provides different MDR. But when the speaker change points obtained from the two sets are combined, the MDR reduces significantly. This is because using MFCC of speech and MFCC of ZFFS different speaker change points are detected, one obtained from vocal tract related speaker information and the other obtained from excitation source related speaker information, respectively. It can be observed from Table I that compared to the individual performance, the MDR reduces for both clean and noisy case in the combination and this is consistent for the two distance metric based methods BIC and KLD. In the current database in case of BIC the combination gives MDR of 15% for clean speech and 18% for noisy speech, whereas in case of KLD the combination gives MDR of 19% for both clean and noisy speech. Thus both gives around 11% improvement compared to the individual performance. This proves that the MFCC of speech and MFCC of ZFFS shows complementary speaker specific information related to vocal tract and excitation source. Typical rate of miss detection in noiseless speech, reported in HUB 4 database [20] is 24.7% [21], in TIMIT database [22] is 15.6% [5] and in SWITCHBOARD database [23] is 29.1% [5]. In the current work it is observed that the combination of MFCC of speech and MFCC of ZFFS feature gives almost equal performance in

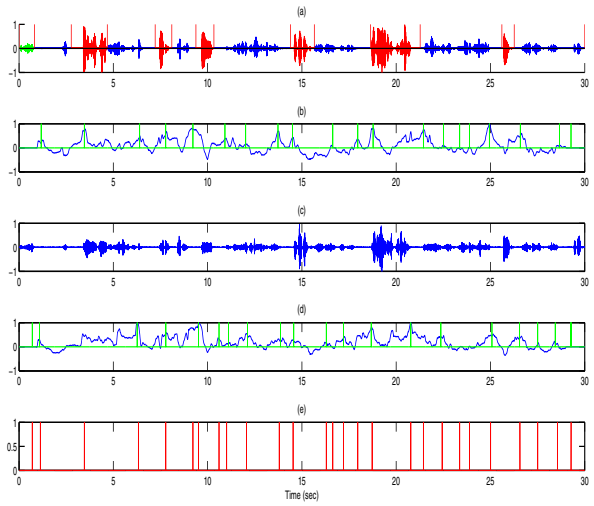

Fig. 3. (a) Speech signal with colour change indicating ground truth marking, (b) Hamming window smoothed BIC distance contour obtained from speech along with speaker change evidences marked in green, (c) ZFFS of speech signal shown in (a), (d) Hamming window smoothed BIC distance contour obtained from ZFFS along with speaker change evidences marked in green, (e) Hypothesized speaker change points obtained from the combination

the clean multivariability speech data as well as in noise mixed speech. Thus the combination provides additional advantage over noisy speech. The experiments performed during the work basically focused on reducing MDR using the advantage of additional source based feature obtained from zero frequency filtered speech signal. Change points given by distance metric based methods related to large dissimilarity between sound units as well along with the speaker changes. As explained in Sec IV-A, the database used in the current work is multi-lingual as well as multi environment. Hence the false detection is observed to be a bit high during the current experiments. Hence our future work will focus on reducing the FAR using speech specific knowledge.

V. Conclusion

Searching speaker specific feature related to excitation source for speaker change detection is the objective of this work. MFCC feature is computed from the excitation source signal obtained by filtering the speech signal using zero frequency filter. Speaker change evidences obtained by this new feature set is combined with the speaker change evidences obtained from MFCC of speech signal. Thus speaker specific information obtained from both source and system is combined and reduced rate of MDR is obtained. The experiments are performed using BIC and KLD based speaker change detection method with both clean and noisy speech signal. It is observed that the proposed feature set also shows robustness to noise. Over all the four cases MDR reduced by average 11.4% whereas the FAR remains almost at the same range. In our future work speech signal specific knowledge will be used to reduce the FAR.

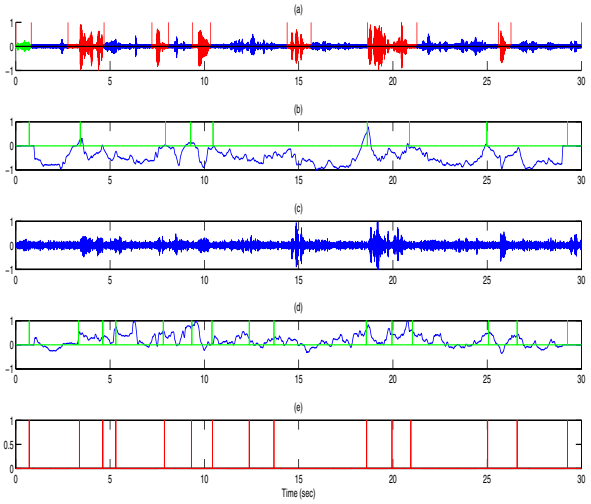

Fig. 4. (a) Noisy speech signal with colour change indicating ground truth marking, (b) Hamming window smoothed BIC distance contour obtained from speech along with speaker change evidences marked in green, (c) ZFFS of noisy speech signal shown in (a), (d) Hamming window smoothed BIC distance contour obtained from ZFFS along with speaker change evidences marked in green, (e) Hypothesized speaker change points obtained from the combination

REFERENCES

[1] M.H. Moattar and M.M. Homayounpour, *A review on speaker diarization systems and approaches*, Speech Communication, Vol. 54, pp. 1065 - 1103, 2002.

[2] X. A. Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, *Speaker Diarization: A Review of Recent Research*, IEEE Transactions on Audio, Speech and Language Processing, Vol. 20, No. 2., pp. 356 - 370, 2012.

[3] M. Kotti, E. Benetos and C. Kotropoulos, *Automatic Speaker Change Detection with the Bayesian Information Criterion using MPEG-7 Features and a Fusion Scheme*, Proc. of IEEE International Symposium on Circuits and Systems, Island of Kos, 2006.

[4] S. S. Chen and P.S. Gopalakrishnan, *Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion*, Proc. of DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[5] P. Delacourt and C.J. Wellekens, *DISTBIC: A speaker-based segmentation for audio data indexing*, Speech Communication, Vol. 32, pp. 111 - 126, 2000.

[6] J. W. Hung, H. Wang and L. Lee, *Automatic metric-based speech segmentation for broadcast news via principal component analysis*, Proc. of INTERSPEECH, pp. 121 - 124., 2000.

[7] L. Lu and H. J. Zhang, *Unsupervised speaker segmentation and tracking in real-time audio content analysis*, Multimedia Systems, Vol. 10, Issue 4, pp. 332 - 343, 2005.

[8] C. Barras, X. Zhu, S. Meignier and J.L. Gauvain, *Multistage speaker diarization of broadcast news*, IEEE Transactions on Audio, Speech and Language Processing, Vol.14, No.5, pp. 1505 - 1512, 2006.

[9] M.A.Siegler, U. Jain, B. Raj and R.M. Stern, *Automatic segmentation, classification and clustering of broadcast news audio*, Proc. of DARPA Speech Recognition Workshop, Chantilly, pp. 97 - 99, 1997.

[10] T. Wu, L. Lu, K. Chen and H. Zhang, *Universal background models for real-time speaker change detection*, Proc. of the 9th International Conference on Multimedia Modeling, Tamshui, Taiwan, pp. 135 - 149, 2003.

[11] T. Liu, X. Liu and Y. Yan, *Speaker Diarization SystemBased on GMM and BIC*, International Symposium on Chinese Spoken Language Processing, Singapore, 2006.

[12] H. S. Beigi and S. Maes, *Speaker, Channel and Environment Change Detection*, In: World Congress of Automation. Proc. of World Congress of Automation, 1998.

[13] L. Lu and H. Zhang, *Speaker change detection and tracking in real-time news broadcast analysis*, Proc. of the ACM Multimedia, France, pp. 602 - 610, 2002.

[14] C.H. Wu and C.H. Hsieh, *Multiple change-point audio segmentation and classification using an MDL-based Gaussian model*, IEEE Trans Audio Speech Language Processing. Vol. 14, No. 2, pp. 647 - 657, 2006.

[15] S. Meignier, D. Moraru, C. Fredouille, J.F. Bonastre and L. Besacier, *Step-by-step and integrated approaches in broadcast news speaker diarization*. Computer Speech and Language, Vol. 20, Issues. 2 - 3, pp. 303 - 330, 2006.

[16] L. Lu and H.J. Zhang, *Real-time unsupervised speaker change detection*, Proc. of ICPR, Vol. 2, Quebec City, Canada, 2002.

[17] S.M. Chu, H. Tang and T.S. Huang, *Fishervoice and semi- supervised speaker clustering*, Proc. of ICASSP, pp. 4089 - 4092, 2009.

[18] Haris B. C. , G. Pradhan, A . Misra, S.R.M. Prasanna, R.K. Das and R. Sinha, *Multivariability speaker recognition database in Indian scenario. International Journal of Speech Technology*, Vol.15 pp. 441 - 453, 2012.

[19] K. S. R. Murty and B. Yegnanarayana, *Epoch Extraction From Speech Signals*, IEEE Transactions on Audio, Speech and Language Processing, Vol. 16, No. 8, pp. 1602 - 1613, 2008.

[20] J. Alabiso, R. MacIntyre and D. Graff, *1997 English Broadcast News Transcripts (HUB4)*, Linguistic Data Consortium, Philadelphia, 1998.

[21] A. Tritschler and R. Gopinath, *Improved speaker segmentation and segments clustering using the Bayesian information criterion*, Proc. 6*th* European Conference Speech Communication and Techology, pp. 679-682, Budapest, Hungary, September, 1999.

[22] J. S. Garofolo, *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, Philadelphia, 1993.

[23] J. J. Godfrey and E. Holliman, *Switchboard-1 Release 2*, Linguistic Data Consortium, Philadelphia, 1997.