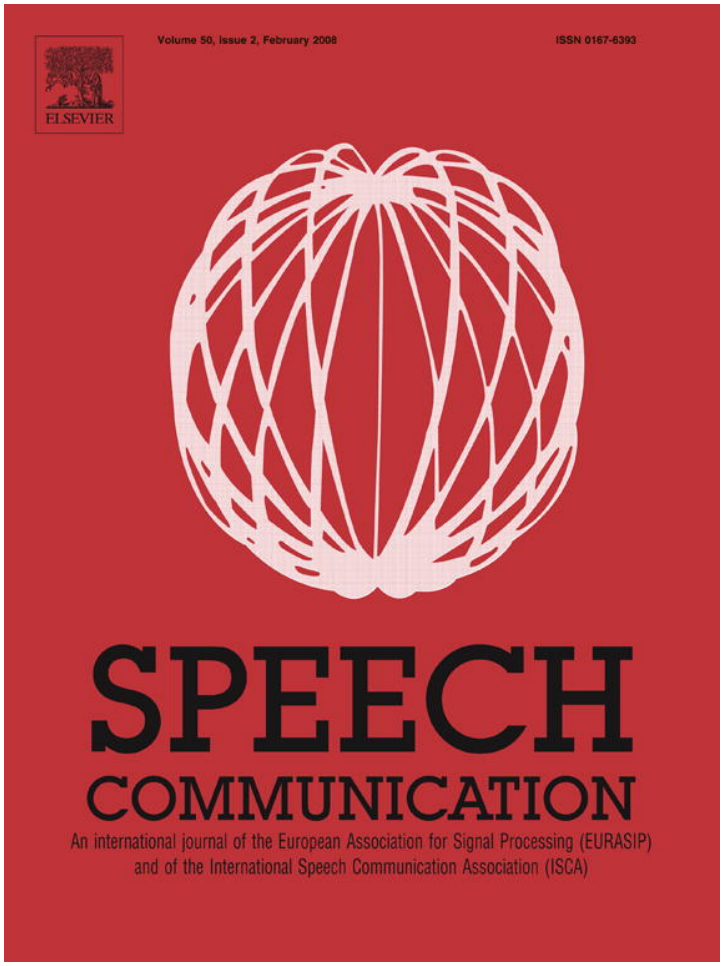

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 50 (2008) 153–161
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Speaker change detection in casual conversations using excitation source features

N. Dhananjaya [a,*], B. Yegnanarayana [b]

[a] Indian Institute of Technology Madras, Chennai, India
[b] International Institute of Information Technology, Hyderabad, India

Received 27 June 2006; received in revised form 26 July 2007; accepted 20 August 2007

Abstract

In this paper we propose a method for speaker change detection using features of excitation source of the speech production mechanism. The method uses neural network models to capture the speaker-specific information from a signal that represents predominantly the excitation source. The focus in this paper is on speaker change detection in casual telephone conversations, in which short (<5 s) speaker turns are common. Excitation source features are a better choice for modeling a speaker, when limited amount of speech data is available, when compared to the vocal tract system features. Linear prediction residual is used as an estimate of the excitation source signal. Autoassociative neural network models are proposed to capture the higher order relations among the samples of the residual signal. Speaker models are generated for every one second of voiced speech from the first few seconds of the conversation. These models are used to detect the speaker change points. Performance of the proposed method for speaker change detection is evaluated on a database containing several two-speaker conversations.

Keywords: Speaker change detection; Multispeaker conversation; Autoassociative neural network (AANN) models; Excitation source features; Linear prediction (LP) residual

1. Introduction

Speaker change detection is the task of identifying the speaker change points in the monaural (single microphone) recording of a conversation between two or more speakers. A block schematic of the standard approach for speaker change detection is shown in Fig. 1. At every instant, a dissimilarity score is computed between features extracted from two blocks of data on either side of the instant. Most methods tend to use a statistical dissimilarity measure on the features representing the vocal tract system (Gish et al., 1991; Chen and Gopalakrishna, 1998; Johnson, 1997; Delacourt and Wellekens, 2000; Makhoul et al., 2000; Lu and Zhang, 2002). Performance of speaker change detection based on statistical methods may degrade, when the speaker turns in the conversation are short, *i.e.*, <5 s. The reason is that there is not enough data to obtain reliable estimation of the statistical parameters (used in the computation of the dissimilarity) from each block. Moreover, the initial estimates of the speaker change points are refined by building a model for each of the speakers from one or more segments of speech that are most similar. This would require large amount of speech data for a given speaker, and that the regions of that data have been identified in the first step itself. In the absence of such large amount of speech data for a given speaker, and with short speaker turns, it is necessary to look for methods other than those based on statistical dissimilarity of features. In this paper we propose a method for speaker change detection using excitation source features, which does not require long (>5 s) speaker turns. The most popular application of speaker change detection has been

* Corresponding author. Tel.: +91 44 2257 5382; fax: +91 44 2257 4352.
E-mail addresses: dhanu@cs.iitm.ernet.in (N. Dhananjaya), yegna@iiit.ac.in (B. Yegnanarayana).

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.08.003

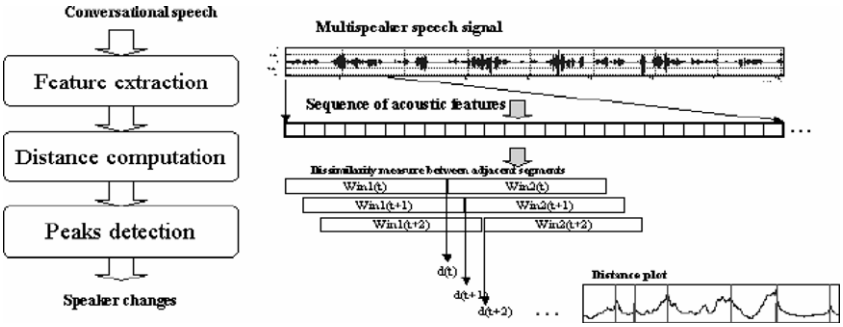

Fig. 1. Block schematic of a typical approach for speaker change detection.

in the indexing of broadcast news data, due to the availability of long speaker turns. The proposed method is useful for speaker change detection in casual conversations and may be useful in forensic applications. The proposed method may also provide complementary or supplementary evidence for speaker change detection, to the evidence already provided by the use of vocal tract features.

The paper is organized as follows: In Section 2 we discuss the statistics of the durations of the speaker turns in casual conversations, to justify the need for exploring methods for speaker change detection in multispeaker data with short speaker turns. The details of the database used for experiments in this paper, is also given in this section. Development of neural network models for capturing speaker-specific information from the excitation source signal is discussed in Section 3. The proposed method for speaker change detection is presented in Section 4. Performance of the proposed method for speaker change detection on some standard databases is discussed in Section 5. In Section 6 the performance of the proposed method is compared with the performance of one of the standard methods based on vocal tract features. Section 7 gives a summary and a few pointers for further exploration of using excitation source features for speaker change detection.

2. Distribution of durations of speaker turns

The database used for experiments in this paper consists of 30 natural conversations between different pairs of speakers over a telephone, ten each for the male–male, female–female and male–female cases. The conversations form part of the Switchboard-2 Phase III database (Graff et al., 2002) of the linguistic data consortium (LDC) used in NIST-2003 extended task speaker recognition evaluation. Each conversation is of 5 min duration, with both sides of the conversation recorded separately, and stored as a 8 kHz, 8-bit μ-law, stereo signal. The separate recording of speech from the individual speakers enables manual marking of the speaker boundaries. The single-channel two-speaker conversation is obtained by summing the two sides of the conversation. The dataset consists of a total number of 2789 speaker change points in about one hour of voiced speech (63 min) from a total 2.5 h of the conversational speech.

Distribution of the durations of speaker turns varies with the type of multispeaker data. Broadcast news data typically has longer speaker turns compared to the durations of speaker turns in conversational (telephone) speech. The distribution of the durations of the speaker turns for the telephone conversation database described above is shown in Fig. 2. The speaker turn durations are computed only for voiced regions of the speech. The nonvoiced regions (silence, noise and unvoiced regions of speech) in the conversation are detected and eliminated using simple thresholds on the short term energy (STE) of the signal and on the ratio of the corresponding STEs in the signal and its 12th order linear prediction (LP) residual. The size of the short term window used for LP analysis is 20 ms. Using this procedure about 42% of the conversational speech was marked as voiced.

The distribution in Fig. 2 shows that over 60% of the speaker turns have less than 1 s duration. For detection of a speaker change, blocks of speech data on either side of the hypothesized change point are considered. Detection of the change point is affected even if one of the blocks in the pair has short duration, if the change point detection is based on the statistics of the features in the block. It is interesting to note that for this data in over 90% of the cases at least one of the blocks in the pair around the

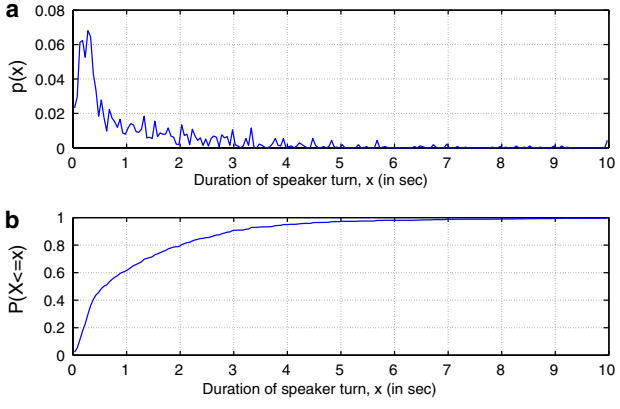

Fig. 2. (a) Frequency distribution of durations of the speaker turns. (b) Cumulative distribution of durations of the speaker turns.

change point has duration less than 1 s. Thus one can see the importance of short duration turns, and the need to develop methods to determine them automatically in conversational speech.

3. Neural network models for extracting speaker-specific features from excitation source signal

In this study we use linear prediction (LP) residual as an approximate representation of the signal exciting the vocal tract system. The residual is computed using a 12th order LP analysis for every 20 ms segments, with a shift of 5 ms. Since the second order statistics representing mostly the resonances of the vocal tract and the glottal roll-off are removed, the residual signal represents mostly the unpredictable part due to excitation. The excitation part may contain significant higher order information in the sequence of samples of the LP residual signal. It is hypothesized that this higher order information may contain significant speaker-specific knowledge. Simple statistical analysis of the LP residual may not bring out the desired speaker-specific knowledge. Hence we propose the use of autoassociative neural network (AANN) models to capture this knowledge automatically from the LP residual signal.

Fig. 3 shows the structure of the AANN model used in these studies. It consists of five layers, represented by P_1L P_2N P_3N P_4N P_5L, where $P_1, P_2, \ldots, P_5$ represent the number of units in different layers, and L and N indicate whether the units in those layers are linear or nonlinear, respectively. In the five layer network $P_1 = P_5$, $P_2 = P_4$ and P_3 is the middle bottleneck layer. The size of the block of samples used for training the model decides the value of P_1. Typically $P_2 > P_1$ and $P_3 < P_1$. The details of the network and its ability to capture the higher order relations among the input samples are described in (Prasanna et al., 2006; Yegnanarayana et al., 2001).

In the present study, an AANN model is trained using blocks of LP residual samples, each block shifted by one sample. The input and the desired output of the network is the same block. The network structure is 40 L 60 N 12 N 60 N 40 L. The structure is estimated over a few trials with different structures (number of units in different layers), and observing the training error as a function of number of cycles or iterations of presentation of data. For a given size of the input block, the size of the network is not critical, in the sense that any choice in a wide range for the number of units in the hidden layer seem to give similar performance (Prasanna et al., 2006).

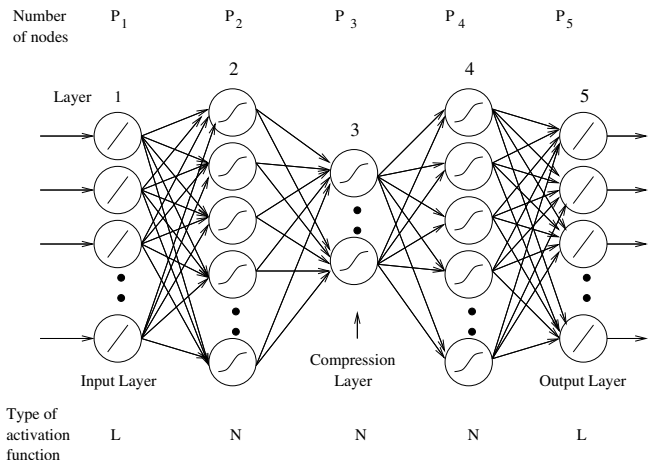

Fig. 3. A five layer autoassociative neural network of structure.

The size of the block in number of samples is determined by the average pitch period of the voiced speech. The size should be as large as possible for the network to capture the higher order relations among the sequence of samples, but should be smaller than the pitch period to avoid the effects due to pitch. We have used a block of 40 samples (chosen empirically), corresponding to 5 ms of the LP residual, and each block is shifted by one sample. Each block of data is normalized to unit norm by dividing each sample value in the block with the root-mean-squared value of the samples in the block. The mean-squared error between input and output of the network is computed for a given set of weights for the network, initialized randomly to start with. The weights are adjusted using the standard backpropagation learning law. The adjustment of weights (*i.e.*, training) is carried out for 500 cycles of presentation of data, where each cycle consists presentation of all training data once.

When the LP residual signal for any speech data, is presented to the trained network, the error as a function of time is not uniform across time. Instead of error, one can compute the confidence score given by $c[n] = \exp(-e[n])$, where $e[n]$ is the mean-squared error for the nth block. Fig. 4 shows the LP residual signal and the confidence score as a function of time for a segment of voiced speech. It can be seen that the confidence score is not uniform. The score is generally higher around the regions of glottal closure (region around the large peak values in the LP residual). This property can be utilized for better training and for testing as well.

The network can be better trained using blocks of samples in the region around the glottal closure. The regions are determined by using the glottal closure instants (GCI). The GCIs are detected by picking the positive zero crossings of the average phase slope function computed from the LP residual as outlined in (Smits and Yegnanarayana, 1995). A window (see Fig. 4c) on either side of the GCI is considered for selecting the blocks of data for training.

4. Speaker change detection

Given a multispeaker conversation, if one of the speakers is modeled, then the confidence score is obtained for blocks of data at each instant of the conversation speech data. The confidence scores for the regions of speech belonging to the modeled speaker will be large, and for other speakers the scores will be small. The difference in

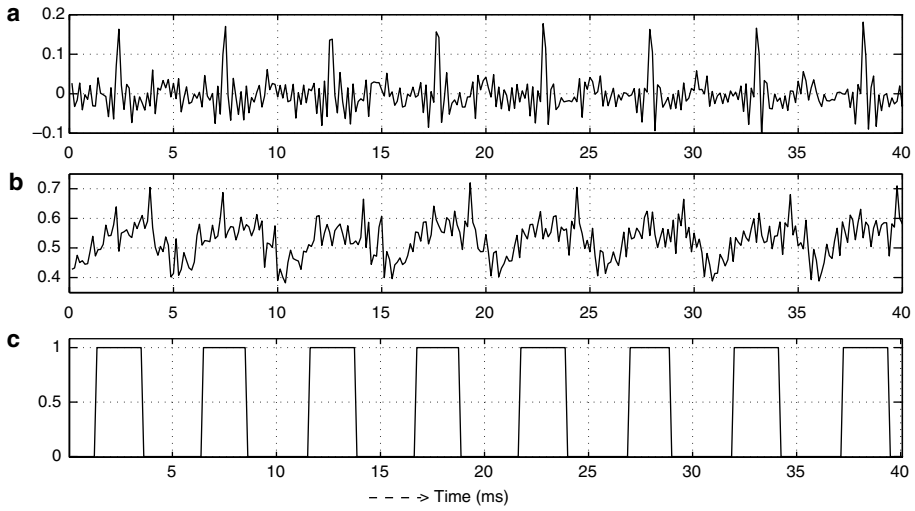

Fig. 4. Frame selection for training AANN models. (a) LP residual of a voiced speech segment. (b) Confidence scores of an AANN model for the input in (a). (c) Weight function defined around the instants of glottal closure.

confidence scores around any instant of time is an indication of a speaker change point. By determining the speaker change points, the segments corresponding to the modeled speaker are removed from the multispeaker data, and another speaker model is developed using the remaining data. The process of building speaker models and removing the corresponding segments in the multispeaker data is repeated until no more speakers are left. Thus the proposed method for speaker change detection has two phases: model generation and change detection.

Two important issues arise in generating a reliable model for the speakers in the conversation. (1) Sufficiency of data for speaker modeling, and (2) automatic detection of single speaker regions.

The amount of data required for developing a speaker model is determined experimentally by using training data of different durations, *i.e.*, 5 s, 2 s, 1 s, 0.5 s and 0.25 s. The single speaker voiced data was manually selected from a two-speaker conversation data. Fig. 5 shows the confidence scores plots (smoothed using moving average over 0.5 s window) obtained by models trained with different durations of single speaker data. From the figure, it appears that about one second of data may be adequate to build a speaker model to discriminate it from other speakers.

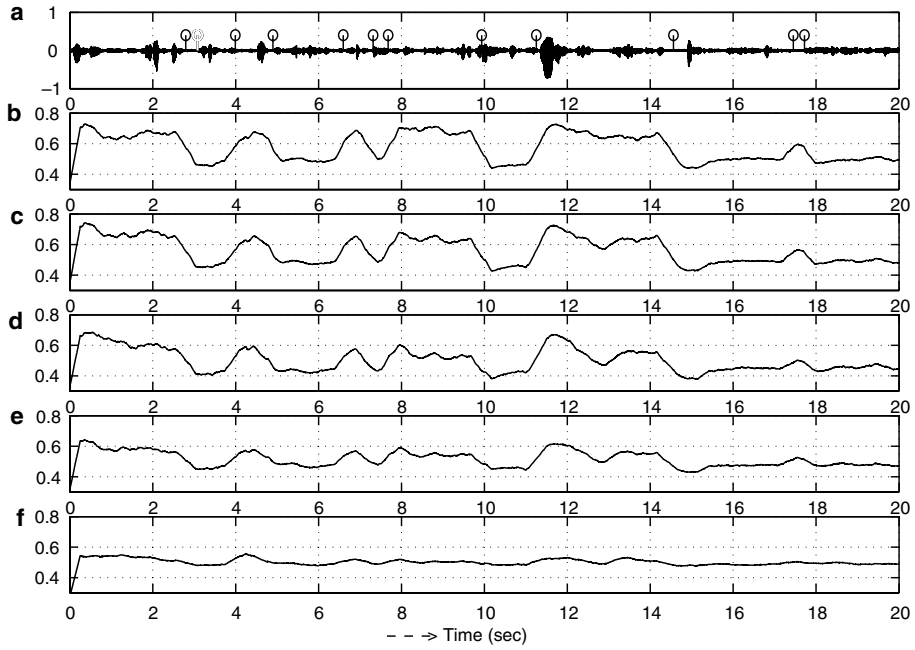

Fig. 5. (a) Waveform of a two-speaker speech signal with the actual speaker changes marked. Evidence (confidence scores) obtained by models built from (b) 5 s, (c) 2 s, (d) 1 s, (e) 0.5 s and (f) 0.25 s of training data.

For all subsequent experiments, one second of training data is used to build speaker models.

In casual conversational speech, it is not guaranteed that a randomly chosen segment of one second voiced speech is due to only one speaker. In order to circumvent this problem, M (about 10) models are built from M adjacent segments (1 s) of speech, with an overlap of half a second between segments. The possibility of at least two pure segments (a pure segment is one which contains only one speaker) is thereby increased. Around five seconds of voiced speech data (after the removal of nonvoiced regions) towards the beginning of the conversation is used to select training data for building the models. The entire conversation is tested against each of the M models to obtain the confidence scores. The similarity between any two models can be measured in terms of the normalized cross-correlation coefficient value between the two mean-subtracted confidence score plots. The confidence plots are smoothed using a moving average window of size 0.5 s, before computing the correlation coefficient values. If the value of the cross-correlation coefficient is high and positive (close to $+1$), then the training data corresponding to these models belong to the same speaker. If the value of the cross-correlation coefficient is high and negative (close to -1), then the training data corresponding to these models belong to different speakers. It is unlikely that any two training data sets containing speech of more than one speaker, will give a high value of correlation coefficient.

The values of cross-correlation coefficient computed between confidence plots of the ten models generated from a two-speaker conversation, are given in Table 1. The highest value in each row is highlighted after ignoring the diagonal value and values corresponding to models trained from adjacent segments. For example, the models M_1, M_8 and M_9 show a good degree of similarity, and hence the corresponding training data sets belong to the same speaker. Similarly, the models M_4 and M_6 have good similarity, and hence the corresponding training data sets belong to the second speaker. Some models belonging to different speakers yield negative correlation values (e.g. M_4 and M_8). Model M_3 seems to have been trained from an impure segment, and hence has relatively lower correlation coefficient values. The pair of models with the highest absolute value of the cross-correlation coefficient (0.61 between M_1 and M_9 from Table 1) are hypothesized to be trained from pure segments of speech and belonging to the same speaker.

The change detection phase has two steps: Combining evidence from multiple models and detection of peaks in the combined sequence of confidence scores. In general, N out of the M models that give high correlation coefficient values with one another can be selected for combining the evidence. In this paper, evidence from two models ($N = 2$) with the highest value of correlation coefficient is used. A difference sequence is computed from the sequence of confidence scores obtained from a model, by using a simple difference operator. The difference operator computes at any instant of time, the difference of average confidence scores between two windows of size T_d seconds, on either side of the time instant. The window size of the difference operator used can vary the performance of the speaker change detection task, and its effect on the performance is studied in Section 5. The two difference sequences corresponding to the two chosen models are combined by simple averaging. A peak in the combined difference sequence corresponds to a significant change in the levels of confidence scores on either side of the peak, and is a likely candidate for a speaker change. The peaks are detected by picking the positive zero crossings of the output of a simple difference operator (window size T_d) on the combined difference sequence. As a first step, all peaks in the combined difference sequence are hypothesized as speaker change points. In the next step, each of these hypothesized speaker change points is validated in order to reduce the number of false alarms. The hypothesis is considered valid, if the peak value is greater than $\lambda = (\mu - \alpha\sigma)$, where μ and σ are the mean and the standard deviation of the peak values in the difference sequence. α is a constant parameter which controls the threshold λ, and hence the performance of the speaker change detection task. Fig. 6 shows the results of the peak detection and validation process on a two-speaker conversation, for a threshold of $\lambda = (\mu - 0.25\sigma)$.

Table 1
Cross-correlation coefficient values between confidence score plots of 10 models generated from adjacent, overlapped segments of a male–male conversational speech data

Models	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}
M_1	1.00	0.56	0.06	−0.18	−0.12	−0.06	0.49	0.54	**0.61**	0.38
M_2	0.56	1.00	0.26	−0.05	−0.00	0.05	0.41	0.38	**0.45**	0.45
M_3	0.06	0.26	1.00	0.54	0.35	**0.38**	0.25	−0.07	−0.01	0.28
M_4	−0.18	−0.05	0.54	1.00	0.62	**0.46**	0.09	−0.31	−0.29	0.02
M_5	−0.12	−0.00	**0.35**	0.62	1.00	0.49	0.09	−0.22	−0.22	0.03
M_6	−0.06	0.05	0.38	**0.46**	0.49	1.00	0.24	−0.19	−0.18	0.15
M_7	0.49	0.41	0.25	0.09	0.09	0.24	1.00	0.36	0.42	**0.50**
M_8	**0.54**	0.38	−0.07	−0.31	−0.22	−0.19	0.36	1.00	0.65	0.30
M_9	**0.61**	0.45	−0.01	−0.29	−0.22	−0.18	0.42	0.65	1.00	0.47
M_{10}	0.38	0.45	0.28	0.02	0.03	0.15	**0.50**	0.30	0.47	1.00

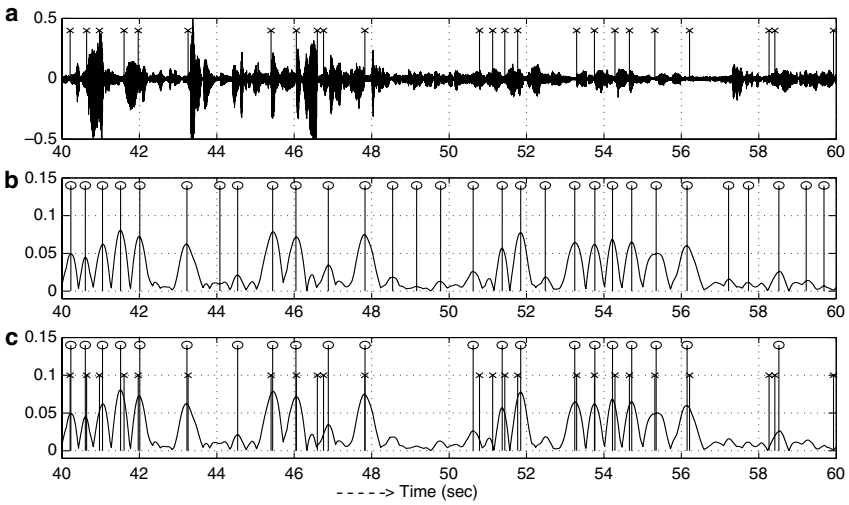

Fig. 6. Speaker change hypothesis and validation on a two-speaker conversation. (a) The speech signal with actual speaker changes marked manually. (b) The combined difference plot with the hypothesized speaker changes. (c) The final validated speaker changes. The actual and detected speaker changes are marked by crosses and circles, respectively.

It can be seen from the figure that the validation eliminates several spurious peaks, thereby reducing the number of false hypotheses of speaker change points.

5. Performance of the speaker change detection task

The performance of the speaker change detection task is evaluated on the database described in Section 2. The performance metrics used and the experimental results, are discussed in this section.

The hypothesized speaker change point is said to be correct if it lies within a tolerance limit from the manually marked speaker change point. Setting a fixed tolerance limit may not be appropriate due to varying durations of the speaker turns. A variable tolerance limit of half the duration of smaller of the two adjacent speaker turns, and not exceeding an upper limit of 0.25 s, is used. The performance of the speaker change detection task is evaluated using the false alarm rate (FAR) and the missed detection rate (MDR). These are defined as follows:

$$P_{\mathrm{FAR}} = \frac{N_{\mathrm{FA}}}{N_{\mathrm{HYP}}} \times 100\%, \tag{1}$$

and

$$P_{\mathrm{MDR}} = \frac{N_{\mathrm{MISS}}}{N_{\mathrm{ACT}}} \times 100\%, \tag{2}$$

where N_{ACT} is the # actual speaker change points (manually marked), N_{HYP} is the # speaker change points hypothesized, N_{FA} is the # false alarms (a hypothesized change point being wrong), N_{MISS} is the # actual speaker change points missed out.

An ideal system should give an FAR of 0% and an MDR of 0%.

Two parameters that affect the performance of the speaker change detection task are: (1) The window size T_{d} of the difference operator used to compute the change in confidence scores, and (2) the threshold λ used for validating the initial speaker change hypotheses. The performance of the proposed method for speaker change detection is given in Table 2, for varying window sizes T_{d} of the difference operator. It is seen that, an increase in the window size reduces the false alarms significantly, but with a slight increase in the missed detections. The performance of the proposed method for speaker change detection for varying threshold parameter is given in Table 3. It can be seen that the choice of the threshold is a trade off between a low false alarm rate and a low missed detection rate.

Speaker change detection results in a sequence of segments, each of which is hypothesized to contain the data of a single speaker. Speaker segregation is a task closely associated with speaker change detection, which involves

Table 2
Performance of the speaker change detection task for varying window size T_{d} of the difference operator

T_{d} (s)	P_{FAR} (%)	P_{MDR} (%)
0.1	45.9	15.7
0.25	33.1	19.1
0.5	22.3	25.9

A peak validation threshold of $\lambda = \mu - 0.25\sigma$ was used.

Table 3
Performance of the speaker change detection task for different values of the threshold λ

Validation threshold λ	P_{FAR} (%)	P_{MDR} (%)
No validation	41.4	16.0
$\mu - 0.5 * \sigma$	35.1	19.3
$\mu - 0.25 * \sigma$	22.3	25.9
μ	18.1	33.7

Size of the window in the difference operator is $T_{\mathrm{d}} = 0.5$ s.

assigning speaker labels to each of these segments. An agglomerative clustering is performed to group the segments into two groups. The algorithm starts with as many number of clusters as there are segments. The number of clusters is reduced iteratively, by combining two segments at a time which are most similar. Absolute difference between the average confidence scores of two segments is used for similarity measure. The clustering process is continued until only two groups are left. The two groups are now hypothesized to represent each of the speakers involved in the conversation. The results of the speaker segregation process are shown in Fig. 7c.

The performance of the speaker segregation task is measured in terms of the *segregation cost* function given by

$$C_{\text{seg}} = 1 - T_{\text{c}}/T_{\text{t}}, \tag{3}$$

where T_{c} is the duration of voiced speech that is correctly segmented and T_{t} is the total duration of the voiced speech in the conversation. The cost function defined here is similar to the one defined in the NIST-2002 speaker recognition evaluation (SRE) plan (Martin and Przybocki, 2002), except that the errors due to speech–nonspeech segmentation is not included here. The segregated speakers need to be mapped to the actual speakers (manually marked) in the conversation, before computing the segregation cost. In order to find out the best mapping, T_{c}, the amount of conversation data segregated correctly, is computed for all possible mappings between the identified speakers and the actual speakers. The mapping for which the best value of T_{c} is obtained, is chosen for computing the segregation cost. The cost function is normalized by a factor C_{def}, which is the minimum segregation cost that can be obtained even without processing the conversation (by assigning the entire conversation to either of the speakers). The default segregation cost C_{def} and the normalized segregation cost C_{norm} are given by

$$C_{\text{def}} = \min(C_{\text{seg}}|\text{All speech labeled as spk 1}, \; C_{\text{seg}}|\text{All speech labeled as spk 2}), \tag{4}$$

and

$$C_{\text{norm}} = C_{\text{seg}}/C_{\text{def}}. \tag{5}$$

A good system should give a C_{norm} value close to zero, and a value close to one is as good as not processing the conversation.

The speaker segregation task prefers a few additional false alarms rather than missing out a genuine speaker change. In other words, oversegmentation is preferred at the speaker change detection stage. Hence no validation of the peaks is performed during speaker change detection. The performance of the speaker segregation task for varying sizes of the difference operator window is given in Table 4. It can be seen that the window size of 0.1 s gives slightly better performance due to oversegmentation and a low miss rate. The size of the difference operator window shows little effect on the speaker segregation performance when measured using the segregation cost. This is mainly because the durations of short speaker turns, which may not get detected with larger window size, do not contribute much to the segregation cost. The C_{seg} values show that around

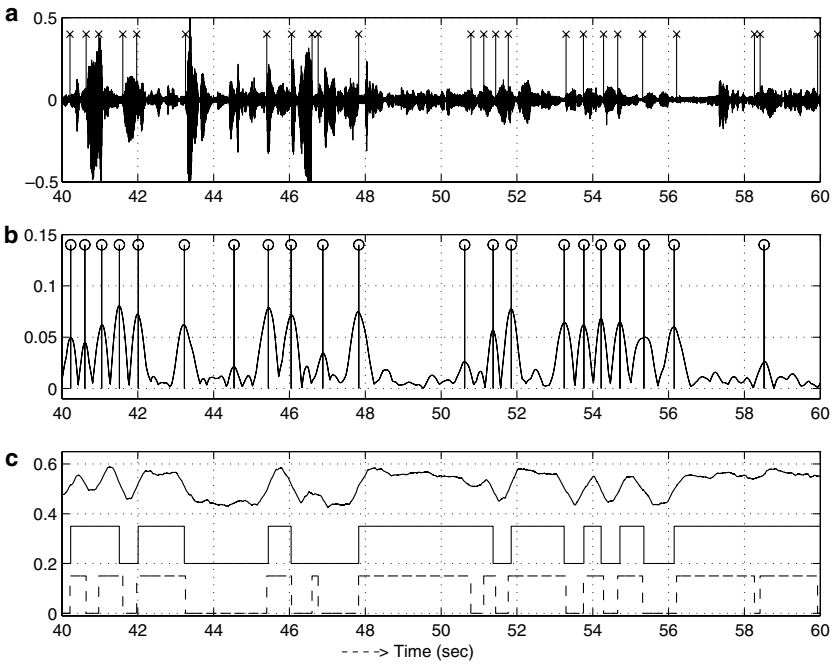

Fig. 7. (a) Two-speaker speech signal with the manually marked speaker change points. (b) The combined difference plot with the hypothesized speaker change points marked. (c) Results of the speaker segregation task. Top: combined (average) and smoothed (0.5 s moving average window) confidence score plot. Middle: a binary signal (solid line) indicating the two speakers as labeled by the speaker segregation task. Bottom: the desired speaker segregation (dashed plot) as per the manual marking.

Table 4
Performance of the speaker segregation task for varying sizes of the difference operator window size T_d

T_d (s)	C_{seg} (%)	C_{norm} (%)
0.1	6.20	16.7
0.25	6.71	18.1
0.5	7.89	21.3

94% of the overall speech is assigned to the correct speaker. However, a C_{def} value of 37% signifies that around 63% of speech (contribution of the dominant speaker) in the dataset can be correctly labeled even without processing the conversations. Hence, using C_{def} as a reference, the C_{norm} values show an accuracy of around 83%.

6. Comparison of the proposed method with a standard method using vocal tract features

A distance-metric based method for speaker change detection using vocal tract features, followed by a model-based resegmentation (Chan et al., 2006), is implemented for comparison with the proposed method which uses excitation source features. The feature vector used consists of 12 mel frequency cepstral coefficients (MFCCs) and log energy computed for every 20 ms frame, shifted by 10 ms. The speaker change detection uses the two-stage DISTBIC technique proposed by Delacourt (Delacourt and Wellekens, 2000). The first stage uses the Kullback–Leibler (KL) distance to hypothesize speaker change points, which are validated in the second stage using the delta Bayesian information criterion (ΔBIC) as dissimilarity measure. The window size used in both the stages is two seconds. A hierarchical agglomerative clustering is performed on the segments obtained after speaker change detection, till only two clusters are left. The ΔBIC is used again as the distance measure for clustering. A 32 mixture Gaussian mixture model (GMM) is developed for each of the two clusters. These two GMM models are used to perform a Viterbi resegmentation.

The performance of the method using the vocal tract features is compared with the proposed method using excitation source features, and the results are given in Table 5. The difference operator window size of 0.5 s and a peak validation threshold of $\lambda = \mu - 0.25 * \sigma$ are used for the SCD results using the excitation source features. The C_{norm} value given in the table is for a window size of 0.1 s. It is seen that the proposed method for speaker change detection using the excitation source features performs significantly better than the method using vocal tract features. Also, the speaker segregation performance of the proposed method is better than the method using the vocal tract features.

Table 5
Performance of the speaker change detection and segregation tasks using different feature type

Feature type	P_{FAR} (%)	P_{MDR} (%)	C_{norm} (%)
Excitation source	22.3	25.9	16.7
Vocal tract	38.4	35.2	43.4

7. Summary and conclusions

The standard methods for speaker change detection provide a statistical solution to detect a point phenomenon (speaker change). They rely on the vocal tract features which require large amounts of data for collecting reliable statistics for detecting speaker change points or for building speaker models. This limits the use of these standard methods for speaker change detection in casual conversations that contain a large number of short (<5 s) speaker turns. An alternate method for detecting the speaker change points was proposed in this paper. The method uses features based on the excitation source information. AANN models were used to capture the speaker-specific information in the LP residual signal. The proposed method for speaker change detection in casual conversations was found to perform better than the method using the vocal tract features.

Some of the directions for continuing this research are as follows: The proposed method requires around 5 s of voiced data towards the beginning of the conversation as adaptation data. Some adaptation time is required as well for the generation of models. But once the models are generated, the conversations can be segmented on a real-time basis. The standard methods for speaker change detection typically involve multiple passes and require a large amount of adaptation data. Hence the proposed method can be a better choice for real-time processing of conversational speech, if the training time of the models can be optimized. It can be observed from Fig. 5 that models trained with data less than one second also contain some evidence for detecting the speaker change points. Reducing the amount of training data used to build models, increasing the number of models used for combining evidence, and the use of efficient methods to combine evidence from multiple models may help reduce the overhead involved in training the models. The current work was focused on detecting speaker changes in casual telephone conversations. The performance of the proposed algorithm has to be studied for different types of data, with different channel and noise considerations. The high values of FAR and MDR underline the difficulty in identifying speaker change points due to short speaker turns.

References

Gish, H., Siu, M., Rohlicek, R., 1991. Segregation of speakers for speech recognition and speaker identification. In: Proc. Internat. Conf. on Acoustics Speech and Signal Processing, Vol. 2, Toronto, Canada, pp. 873–876.

Chen, S., Gopalakrishna, P., 1998. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. Proc. DARPA Broadcast News Transcription and Understanding Workshop. Morgan Kaufmann, San Mateo, CA, pp. 127–132.

Johnson, S., 1997. Speaker Tracking. Master's Thesis. Cambridge University Engineering Department, UK.

Delacourt, P., Wellekens, C.J., 2000. DISTBIC: a speaker-based segmentation for audio data indexing. Speech Comm. 32, 111–126.

Makhoul, J., Kubala, F., Leek, T., Liu, D., Nguyen, L., Schwartz, R., Srivastava, A., 2000. Speech and language technologies for audio indexing and retrieval. Proc. IEEE 88 (8), 1338–1353.

Lu, L., Zhang, H., 2002. Speaker change detection and tracking in real-time news broadcasting analysis. In: Proc. 10th ACM Multimedia. Juan-les-pins, France, pp. 602–610.

Graff, D., Miller, D., Walker, K., 2002. Switchboard-2 Phase III Audio. Linguistic Data Consortium, Philadelphia, USA.

Prasanna, S.R.M., Gupta, C.S., Yegnanarayana, B., 2006. Extraction of speaker-specific excitation information from linear prediction residual of speech. Speech Comm. 48 (10), 1243–1261.

Yegnanarayana, B., Reddy, K.S., Kishore, S.P., 2001. Source and system features for speaker recognition using AANN models. In: Proc. Internat. Conf. on Acoustics Speech and Signal Processing, Vol. 1. Salt Lake City, Utah, USA, pp. 409–412.

Smits, R., Yegnanarayana, B., 1995. Determination of instants of significant excitation in speech using group delay function. IEEE Trans. Speech Audio Process. 3 (5), 325–333.

Martin, A., Przybocki, M., 2002. The NIST Speaker Recognition Evaluation Plan. National Institute of Standards and Technology, USA. <http://www.nist.gov/speech/tests/spk>.

Chan, W.N., Lee, T., Zheng, N., Ouyang, H., 2006. Use of vocal source features in speaker segmentation. In: Proc. Internat. Conf. on Acoustics Speech and Signal Processing, Toulouse, France, pp. 657–660.