Sādhanā Vol. 34, Part 5, October 2009, pp. 729–754. © Indian Academy of Sciences

Application of combined temporal and spectral processing methods for speaker recognition under noisy, reverberant or multi-speaker environments

P KRISHNAMOORTHY and S R MAHADEVA PRASANNA

Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, Guwahati 781 039
e-mail: pkm@iitg.ernet.in,prasanna@iitg.ernet.in

MS received 10 September 2008; revised 10 March 2009

Abstract. This paper presents an experimental evaluation of the combined temporal and spectral processing methods for speaker recognition task under noise, reverberation or multi-speaker environments. Automatic speaker recognition system gives good performance in controlled environments. Speech recorded in real environments by distant microphones is degraded by factors like background noise, reverberation and interfering speakers. This degradation strongly affects the performance of the speaker recognition system. Combined temporal and spectral processing (TSP) methods proposed in our earlier study are used for pre-processing to improve the speaker-specific features and hence the speaker recognition performance. Different types of degradation like background noise, reverberation and interfering speaker are considered for evaluation. The evaluation is carried out for the individual temporal processing, spectral processing and the combined TSP method. The experimental results show that the combined TSP methods give relatively higher recognition performance compared to either temporal or spectral processing alone.

Keywords. Speaker recognition; speech enhancement; temporal and spectral processing; noisy speech; reverberant speech and multi-speaker speech.

1. Introduction

Speaker recognition is the process of automatically recognizing the speakers on the basis of individuality information from the speech signals. It can be classified into speaker identification and speaker verification. Speaker identification is the process of identifying a speaker from a group of registered speakers. The task of a speaker verification system is to authenticate the claim of a speaker based on the test speech (Campbell 1997). Generally, speaker recognition by a machine involves three stages. They are, (i) feature extraction, (ii) training or modelling and (iii) testing stage. The feature extraction module estimates a set of features from the speech signal that represent speaker-specific information. Training phase creates a speaker-specific model using the extracted features of each speaker. In the testing stage, measure of similarity or distance between unknown speaker and stored models is calculated. The model that has the best score is chosen as the identified speaker. Depending on the text used for training and testing, speaker recognition is further divided into text-dependent and text-independent methods. In text-dependent system same text is used for both training and testing, whereas text-independent systems do not rely on the text being spoken.

Most of the speaker recognition systems are developed using clean speech training data. In practical scenario clean speech is often distorted by different types of degradation like background noise, reverberation and speech from other speakers. Because of this distortion, the speaker-specific features are distorted and therefore there is a mismatch between the trained models and the test speech to be recognized. This mismatch causes degradation in the speaker recognition performance (Ming *et al* 2007). Various methods have been proposed in the literature to improve the robustness of a speaker recognizer to overcome this mismatch problem. Generally, this can be accomplished in different stages as follows.

1.1 *Robustness at the signal level (speech enhancement)*

In this approach degraded speech signals are enhanced before the feature extraction stage. Accordingly, before being transformed into feature vectors, the degraded speech undergoes an enhancement step which tries to filter out the degradation (Ortega-Garcia & Gonzalez-Rodriguez 1996).

1.2 *Robustness at the feature level (feature compensation)*

The features representing the speech signal are designed in order to be less sensitive to the degraded conditions. This is achieved by analysing the influence of the degradation on the speech signal and deriving feature extraction methods that reduce the influence of the degradation (Zilovic *et al* 1998; Heck 2000).

1.3 *Robustness at the classifier level (model compensation)*

The aim of the model compensation approach is to determine the influence of the degradation on the distribution of the speech features and to modify the models used in the recognition to take into account about the influence of the degradation. The robustness can be achieved by integrating a model for the speech signal distortion into the overall classifier model (Ming *et al* 2007) or mapping the classifier model obtained during training to better fit the testing condition (Sankar & Lee 1996).

This work aims to provide the robustness at the signal level using our proposed combined temporal and spectral processing (TSP) methods as a pre-processing stage. In the combined TSP method temporal processing refers to the processing of excitation source information in the time domain. Spectral processing refers to processing the degraded speech in the frequency domain representation of speech. In our earlier study (Krishnamoorthy & Prasanna 2008, 2009) the performance of the proposed TSP methods is evaluated using various objective quality measures depending on the nature of the degradation. The results showed that the combined TSP methods give relatively higher performance than the individual temporal or spectral processing alone. In this work, the main objective is to study the speaker recognition performance using the combined TSP methods as a pre-processing stage. Text-independent speaker identification task is taken for demonstration. The rest of the paper is organized as follows: Section 3 briefly reviews combined temporal and spectral processing of degraded speech. Section 3 explains the database and experimental set-up used in the present study.

Section 4 describes the experimental results for the case of noisy, reverberant and multi-speaker speech. Finally, the summary of the work is presented in Section 5.

2. Combined temporal and spectral processing (TSP) of degraded speech

2.1 *Motivation for the combined temporal and spectral processing method*

Several methods have been proposed in the literature for the enhancement of degraded speech, majority of them can be grouped into spectral processing and temporal processing methods. The spectral processing methods are the most popular techniques for speech enhancement, mainly because of their simplicity and effectiveness. Majority of the spectral processing methods based on the fact that the spectral values of the degraded speech will have both speech and degrading components. The spectral characteristics of degradation are therefore estimated and removed to obtain the enhancement. For example, in case of noisy speech enhancement, most of the spectral processing methods (like spectral subtraction and minimum mean square error (MMSE) short term spectral amplitude (STSA) estimators) first estimate the spectral characteristics of the background noise and derive the gain function for the noisy speech signal to attenuate the noise spectral values (Boll 1979; Kamath & Loizou 2002; Ephraim & Malah 1984). In case of reverberation, the reverberant speech can be divided into two parts namely early reverberation and late reverberation. Early reverberation tends to perceptually reinforce the direct sound and is therefore considered harmless to speech intelligibility, whereas the late ones are deleterious to speech quality and intelligibility (Habets *et al* 2008). In such case, spectral subtraction-based methods estimate the late reverberant spectral density and subtract it from the reverberant speech spectra to obtain the enhanced signal (Lebart & Boucher 2001; Habets *et al* 2008). Similarly, in multi-speaker speech separation the spectral based methods first estimate the pitch of desired and interfering speakers and enhance the pitch and harmonics of the desired speaker (Parsons 1976) and/or attenuate the pitch and harmonics of interfering speaker (Morgan *et al* 1997).

Therefore in the spectral processing approach, both short-term magnitude of degradation and degraded speech spectra are estimated first. According to the suppression rule, a spectral gain function is applied to the magnitude spectra of the degraded speech to obtain enhanced speech spectra. The enhanced magnitude and degraded speech phase spectra are then combined to produce an estimate of clean speech. For time-domain resynthesis, overlap-add (OLA) method is typically used. Simplified block diagram showing the important steps of spectral processing is given in figure 1. In figure, noise estimate block refers to estimate of background noise spectra or late reverberant speech spectra or interfering speaker spectra.

A class of temporal processing methods have been proposed by exploiting the excitation source characteristics of the speech signal for the enhancement (Yegnanarayana *et al* 1999; Yegnanarayana & Satyanarayana Murthy 2000; Yegnanarayana *et al* 2003, 2005). Linear prediction (LP) residual obtained by inverse filtering the speech is used as an estimate of the source of excitation of the vocal tract system (Yegnanarayana *et al* 1999). These temporal processing methods are proposed based on the fact that the significant excitation of the vocal tract takes place at the instants of glottal closure and onset of events like burst, frication and aspiration. The instants of significant excitation will be quasi-periodic in nature during voiced speech and random in nature during unvoiced speech (Murty & Yegnanarayana 2008). Depending on the nature of degradation, the excitation source will have many other random peaks in addition to the original instants of significant excitation. Temporal processing method identifies the original instants of significant excitation and emphasizes the region around them

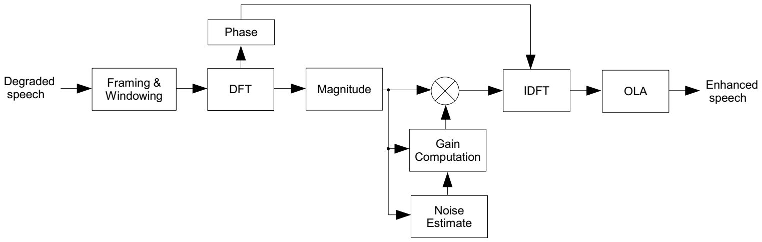

Figure 1. Basic block diagram of the spectral processing approach.

in the excitation source signal to obtain the enhanced speech. The enhancement is achieved by deriving the weight function for the excitation source signal that will emphasize the region around the instant of significant excitation of the original signal and deemphasize the random peaks of degradation. The time-varying all-pole filter derived from the degraded speech is excited by the weighted residual to obtain the temporally processed speech (Yegnanarayana *et al* 1999; Yegnanarayana & Satyanarayana Murthy 2000; Yegnanarayana *et al* 2003, 2005).

Generally, there will be changes in the excitation characteristics both at the fine and gross levels during speech production. The fine level changes may be from closed phase to open phase in a pitch period and the gross level changes may be from silence to voiced excitation. The weight function for the excitation source signal is derived at two different levels, namely, gross and fine levels to obtain the enhanced signal. The gross level weight function is derived to identify the speech and non-speech regions of degraded speech signal and the fine level weight function is derived to enhance the instants of significant excitation of original signal. Figure 2 depicts the steps involved in the conventional excitation source information based temporal processing method.

In summary, the underlying principle of processing degraded speech is different in each domain of processing. Most of the spectral processing is based on the estimation and elimination of degradation. The merit of these methods is the effectiveness in eliminating the degradation, since they are explicitly estimated. The demerit is the need for explicit modelling of spectral characteristics of degradation. This is a difficult task for highly non-stationary environments. On the other hand, excitation source information based temporal processing methods first identify the speech-specific features at the gross and fine levels and enhance those features to obtain the enhancement. Hence an information about the degradation is not

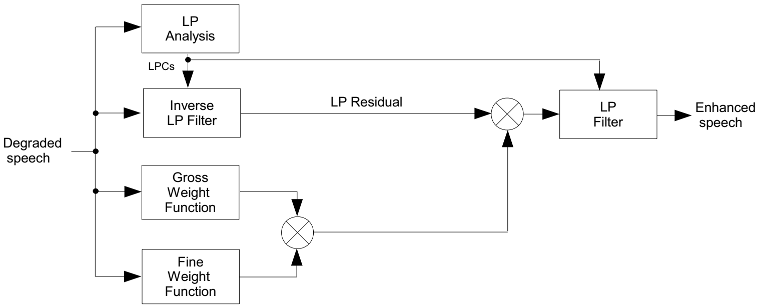

Figure 2. Basic block diagram of the temporal processing approach.

mandatory in the enhancement process. The merit is the effectiveness in the enhancement of speech-specific regions and do not require explicit modelling of degradation. The demerit may be the ineffectiveness in minimizing the degrading component, since it is not explicitly modelled. It may be possible that one domain of processing may aid other domain of processing in minimizing the demerit. For instance,

(i) In noisy speech enhancement, the difficulty in estimating degradation in the highly non-stationary environment for spectral processing may be carried out by the gross weight function derived from the temporal processing as a voice activity detector to identify non-speech regions.

(ii) In reverberant speech enhancement, spectral subtraction-based spectral processing methods reduce the late reverberation (i.e. tail of the impulse response) by estimating and subtracting the late reverberant spectrum from the degraded speech spectrum. On the other hand, these methods may not provide better enhancement in high signal to reverberation ratio (SRR) regions, i.e. early reverberant part of the signal. However, the excitation source information-based temporal processing methods mainly enhance the speech-specific features of high SRR regions in the temporal domain (Krishnamoorthy & Prasanna 2009).

(iii) In multi-speaker case, spectral based speech separation methods depend on the differences in the fundamental frequency characteristics to enhance the spectral features of individual speakers. The main difficulty is the accurate estimation of the fundamental frequency of desired speaker. The spectral processing preceded by temporal processing may improve the accuracy of the pitch estimation of desired speaker. This is mainly because temporal processing enhances the instants of significant excitation of desired speaker relative to other speakers. As a result, pitch specific cues of an interfering speaker(s) will be de-emphasized with reference to the desired speaker.

Therefore, we can effectively combine these two approaches to obtain improved performance compared to either temporal processing or spectral processing alone. In our earlier study we have exploited this property and developed combined TSP methods for the enhancement of noisy speech, reverberant speech and two speaker speech. The temporal and spectral processing methods are used in sequential manner in the combined TSP method.

2.2 *Combined TSP for enhancement of noisy speech*

A combined TSP method is developed for noisy speech enhancement by emphasizing high signal to noise ratio (SNR) regions in the temporal domain, and eliminating the degradation and enhancing the speech-specific components in the spectral domain (Krishnamoorthy & Prasanna 2008). The temporal domain processing is performed at two levels: gross and fine levels. The high SNR regions at the gross level are identified by computing the sum of the ten largest peaks in the discrete Fourier transform (DFT) spectrum, the smoothed Hilbert envelope (HE) of the LP residual and the modulation spectrum values. The fine level features (instants of significant excitation) are identified by performing the following: (i) Sinusoidal analysis of noisy speech (McAulay & Quatieri 1986), (ii) Convolving the HE of the LP residual of the speech obtained from sinusoidal analysis by first order Gaussian differentiator (FOGD) (Prasanna & Subramanian 2005). A weight function is derived by combining the gross and fine level features to emphasize the excitation signal around the instants of significant excitation. In the next step, to improve the spectral characteristics and to provide better noise suppression, the temporally processed speech signal is then subjected to spectral processing. In spectral processing the attenuation of background noise spectra is achieved by conventional multi-band spectral subtraction (Kamath & Loizou 2002) or STSA–MMSE estimator (Ephraim & Malah 1984) approach. The various steps involved in the proposed combined TSP method are given in table 1.

2.3 *Combined TSP for enhancement of reverberant speech*

A combined TSP method is proposed for reverberant speech enhancement that suppresses the effect of early and late reverberations (Krishnamoorthy & Prasanna 2009). First, the late reverberant component is suppressed using spectral processing. An estimate of late reverberant power spectrum is obtained and then subtracted from the power spectrum of the reverberant speech. In the next step, temporal processing is performed to attenuate the early reverberation. The high signal to reverberation ratio (SRR) regions are identified using the ten largest peaks in the DFT spectrum, the smoothed HE of the LP residual and the modulation spectrum values. The fine level features are identified from the band-pass filtered HE of the LP residual. A weight function is derived by combining the gross and fine level features. The LP residual of spectrally processed speech is multiplied by the weight function to attenuate the reverberant peaks. The enhanced residual signal and the vocal tract system characteristics derived from the spectrally processed speech are used for synthesizing enhanced speech. Table 2 describes the various steps involved in the proposed combined TSP method for the enhancement of reverberant speech.

2.4 *Combined TSP for enhancement of multi-speaker speech*

A combined TSP method is developed for separating speech of individual speakers from the mixture of two speaker speech signals, collected over a pair of microphones. The time delay in the arrival of speech of each speaker at a pair of microphones is exploited for speech separation. The mixed speech signals are first subjected to temporal processing. In temporal processing, speech of each speaker is enhanced with respect to the other by relatively emphasizing the speech around the instants of significant excitation of desired speaker by deriving speaker-specific weight function. To further improve the separation, the temporally processed speech is subjected to spectral processing. This involves enhancing the regions around the pitch and harmonic peaks of short time spectra computed from the temporally processed speech. An estimation of the pitch is obtained from the temporally processed speech. Table 3 briefly outlines the different steps involved in the proposed combined TSP method for two speaker separation.

3. Database and experimental description

3.1 *Database*

The speaker recognition studies are carried out on the TIMIT database (Zue *et al* 1990). The TIMIT database contains recordings of 630 speakers (438 male and 192 female) of 8 major dialects of American English, each reading 10 phonetically rich sentences of approximately 3 sec each. The speech was recorded using a high quality microphone in a sound proof booth at a sampling frequency of 16 kHz, with no session interval between recordings. Out of 630 speakers, 100 speakers are randomly selected for forming subset for the study. The common way of using this database is to use the first eight utterances of each speaker for the training and the last two utterances for testing (Reynolds 1995). The speech signals of all these speakers are first downsampled to 8 kHz sampling rate.

Table 1. Combined temporal and spectral processing algorithm for enhancement of noisy speech.

Temporal Processing:
Gross Level Processing

- Compute the sum of the ten largest peaks in the discrete Fourier transform (DFT) magnitude spectrum using frame size of 20 ms and shift of 10 ms (Krishnamoorthy & Prasanna 2008). Mathematically, it is expressed as

$$s_d(l) = \sum_{j=1}^{10} |X(k_j, l)| \tag{1}$$

  where l is the frame index, k_j represents the frequency indexes of ten largest spectral peaks and $X(k, l)$ represents the DFT of a frame of noisy speech and is computed as

$$X(k, l) = \sum_{n=0}^{N-1} x(n) w(n - lR) e^{-\frac{j2\pi nk}{N}} \tag{2}$$

  where $w(n)$ is a Hamming window, N is the number of points used for computing the DFT and R is the frame shift in samples.
- Compute linear prediction (LP) residual of noisy speech using a frame size of 20 ms, shift of 10 ms and 10th order LP analysis.
- Compute the Hilbert envelope (HE) of LP residual. The HE of LP residual $e(n)$ is defined as (Rao *et al* 2007)

$$h_e(n) = \sqrt{e^2(n) + e_h^2(n)} \tag{3}$$

  where $e_h(n)$ is the Hilbert transform of $e(n)$, and is given by

$$e_h(n) = IDFT[E_h(k)] \tag{4}$$

  where

$$E_h(k) = \begin{cases} -jE(k), k = 0, 1, \ldots, \left(\frac{N}{2}\right) - 1 \\ jE(k), k = \left(\frac{N}{2}\right), \left(\frac{N}{2}\right) + 1, \ldots, (N - 1) \end{cases} \tag{5}$$

  where IDFT denotes the Inverse DFT and $E(k)$ is computed as DFT of $e(n)$.
- Smooth the HE of LP residual using 50 ms Hamming window.

$$h_{sm}(n) = h_e(n) * h_{w1}(n) \tag{6}$$

  where $*$ denotes convolution operation and $h_{w1}(n)$ is Hamming window of 50 ms duration.
- Compute the modulation transfer function energies of the noisy speech signal. Mathematically the modulation transfer function energies are expressed as (Prasanna *et al* 2009)

$$m(l) = \sum_{p=1}^{18} \left[\sum_{k=k_1}^{k=k_2} |\hat{X}_p(k, l)|^2 \right] \tag{7}$$

  where l is the frame index, p represents the critical band number, k_1 and k_2 represent frequency index of 4 Hz and 16 Hz, respectively (Greenberg & Kingsbury 1997). $\hat{X}_p(k)$ is the DFT of normalized envelope of p^{th} filter output and is computed as described in Eqn. (2).

Table 1. (Continued).

- Enhance the evidences of high SNR regions of each of the above parameters (i.e. sum of the ten largest peaks in the DFT spectrum, the smoothed HE of the LP residual and the modulation spectrum values) using the first order difference of the evidences obtained (refer to Appendix-A).
- Sum all the enhanced parameters and normalize the sum with respect to maximum value.
- Nonlinearly map the normalized sum values using a sigmoid nonlinear function

$$w_g(n) = \frac{1}{1 + e^{-\lambda(s_i(n) - T)}} \tag{8}$$

where slope parameter $\lambda = 20$ and T equal to average value of the normalized sum $s_i(n)$. This generates a gross weight function.

Fine Level Processing

- Compute the DFT magnitude and phase spectra for the noisy speech using 1024 point DFT.
- Pick the largest 8 peaks in the DFT magnitude spectrum and corresponding phase values and synthesize the speech. If the amplitudes, frequencies, and phases that are estimated for the k^{th} segment are denoted by A_l^k, ω_l^k, and θ_l^k respectively, the synthetic speech signal $\tilde{s}^k(n)$ can be represented as (McAulay & Quatieri 1986)

$$\tilde{s}^k(n) = \sum_{l=1}^{L^k} A_l^k \cos(\omega_l^k n + \theta_l^k) \tag{9}$$

where L^k is the number of sinusoidal components in the frame.
- Compute the HE of LP residual of the signal obtained.
- Emphasize the regions around the instants of significant excitation using the neighborhood information of each sample in the HE. The emphasized HE is computed as

$$h(n) = \frac{h_e^2(n)}{\frac{1}{2M+1}\sum_{p=n-M}^{n+M} h_e(p)} \tag{10}$$

where $M = (2 \times F_s/1000)$, $h(n)$ is the emphasized HE of $h_e(n)$ and F_s is the sampling frequency in Hz.
- Obtain first order Gaussian differentiator (FOGD) given by (Prasanna & Subramanian 2005)

$$g_d(n) = \frac{1}{\sigma\sqrt{2\pi}}\left[e^{-\frac{(n+1)^2}{2\sigma^2}} - e^{-\frac{n^2}{2\sigma^2}}\right], \ 1 \le n \le L_g \tag{11}$$

where Gaussian window of length $L_g = 80$ samples and $\sigma = 8$.
- Convolve the negative of FOGD operator with the mean smoothed HE of the LP residual and determine negative to positive transitions.
- Derive the fine weight function $w_f(n)$ by convolving detected instants with the Hamming window

$$w_f(n) = \left(\sum_{i=1}^{N_i} \delta(n - a_i)\right) * h_w(n) \tag{12}$$

where N_i represents total number of detected instants, a_i is the approximate location of instants and $h_w(n)$ is the Hamming window of 3 ms duration (Krishnamoorthy & Prasanna 2008).

Table 1. (Continued).

Final Weight Function

- Multiply the two weight functions (gross and fine weight functions) to generate the final weight function.
- Multiply the LP residual signal of noisy speech by the final weight function.
- Excite the time-varying all-pole filter (derived from noisy speech) using weighted residual to obtain the temporally processed speech.

Spectral Processing:

- Update the noise magnitude spectrum if 5 consecutive frames are detected as non-speech regions.
- Process the temporally processed speech by any of the conventional spectral processing (e.g. multi-band spectral subtraction (Kamath & Loizou 2002) or MMSE estimator (Ephraim & Malah 1984)) methods.
- Reconstruct the enhanced speech signal using IDFT and overlap-add (OLA) method.

Table 2. Combined temporal and spectral processing algorithm for enhancement of reverberant speech (Krishnamoorthy & Prasanna 2009).

Spectral Processing:

- Estimate the late reverberant spectral variance $\hat{S}_{yl}(l,k)$.

$$\hat{S}_{yl}(l,k) = \gamma\omega(l - N_1) * |Y(l,k)|^2 \tag{13}$$

  where $Y(l,k)$ is the short time Fourier transform of reverberant speech $y(n)$. The symbol * denotes convolution in the time domain and $w(l)$ is a smoothing function. γ specifies the relative strength of the late-impulse components and is set to 0·32 and

$$w(l) = \begin{cases} \frac{l+a}{a^2} e^{\frac{-(l-a)^2}{2a^2}}, & l > -a \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

  where a controls the span of the smoothing function and is set to 5 (Krishnamoorthy & Prasanna 2009; Wu & Wang 2006).
- Compute the *a posteriori* signal to reverberant ratio (SRR) and the *a priori* SRR values. The *a priori SRR* $\xi(l,k)$ is calculated as (decision directed approach) (Ephraim & Malah 1984)

$$\xi(l,k) = \eta \frac{|\hat{S}(l-1,k)|^2}{|Y(l,k)|^2} + (1-\eta)\max\{\gamma(l,k) - 1, 0\}. \tag{15}$$

  The value of η is chosen as 0·98 (Ephraim & Malah 1984) and

$$\gamma(l,k) = \frac{|Y(l,k)|^2}{\hat{S}_{yl}(l,k)} \tag{16}$$

  where the term $\gamma(l,k)$ is interpreted as the *a posteriori SRR*. $\hat{S}(l,k)$ is computed as given in Eqn. (19). The following initial condition is used for the first frame

$$\xi_k(l,k) = \eta + (1-\eta)\max\{\gamma(l,k) - 1, 0\} \tag{17}$$

Table 2. (Continued).

- Determine the gain function $G(l, k)$ for the the spectral subtraction from the estimated SRR values.

$$G(l, k) = 1 - \frac{1}{\sqrt{1 + \xi(l, k)}}. \tag{18}$$

- Multiply the gain function with the reverberant speech spectrum.

$$\hat{S}(l, k) = Y(l, k)G(l, k) \tag{19}$$

- Obtain the spectrally processed speech using IDFT and OLA method.

Temporal Processing:

Gross Level Processing

- Compute LP residual of spectrally processed speech using a frame size of 20 ms, shift of 10 ms and 10^{th} order LP analysis.
- Compute the sum of the ten largest peaks in the DFT magnitude spectrum.
- Compute the HE of LP residual and mean smooth using 50 ms Hamming window.
- Compute the modulation spectrum energies of the spectrally processed speech signal.
- Enhance the high SRR regions of each of the above parameters.
- Sum all the enhanced parameters and normalize the sum with respect to maximum value.
- Nonlinearly map the normalized sum values by using a sigmoid nonlinear function with slope parameter $\lambda = 20$ and T equal to average value of the normalized sum. This generates a gross weight function.

Fine Level Processing

- Band pass filter the LP residual of spectrally processed speech into four subbands whose cut-off frequencies are equally spaced in linear scale.
- Compute the HE of LP residual for each subband.
- Sum all the subband HEs.
- Compute the emphasized HE of the LP residual.
- Obtain FOGD operator from Gaussian window of length $L_g = 80$ samples and $\sigma = 8$.
- Convolve the negative of FOGD operator with the emphasized HE of the LP residual and determine negative to positive transitions.
- Derive the fine weight function by convolving detected instants with the Hamming window of 3 ms duration.

Final Weight Function

- Multiply the two weight functions (gross and fine weight functions) to generate the final weight function.
- Multiply the LP residual signal of noisy speech by the final weight function.
- Excite the time-varying all-pole filter (derived from spectrally processed noisy speech) using weighted residual to obtain the enhanced speech.

3.2 *Feature extraction*

Mel-frequency cepstral coefficients (MFCCs) are the most widely used features for speech and speaker recognitions applications (Reynolds 1994; Picone 1993). MFCCs are estimated based on human perception of critical bandwidths (Picone 1993). The mel-frequency scale has a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. For

Table 3. Combined temporal and spectral processing algorithm for two speaker separation.

Temporal Processing

- Compute the LP residual of mic-1 and mic-2 signals using a frame size of 20 ms, shift of 10 ms and 10th order LP analysis.
- Compute the HE of LP residual.
- Estimate the time delays for each speaker by computing cross-correlation of the HEs using a frame size of 50 ms and shift of 5 ms. In the normalized cross-correlation sequence, the displacement of peak with respect to the center sample is considered as the time delay value (Yegnanarayana *et al* 2005).
- Adjust the HEs using the estimated time delays to produce the coherently adjusted HE for each speaker. That is,

$$h_{s1}(n) = \min(h_1(n), h_2(n - d_1)) \tag{20}$$

$$h_{s2}(n) = \min(h_1(n), h_2(n - d_2)) \tag{21}$$

  where $h_1(n)$ and $h_2(n)$ be the normalized HE sequences of speech signals collected at mic-1 and mic-2, respectively and d_1 and d_2 are the time-delays between the two microphone signals.
- Compute the error function. That is,

$$h_{12}(n) = h_{s1}(n) - h_{s2}(n). \tag{22}$$

Gross Level Processing

- Smooth the absolute values of the error function using 50 ms Hamming window.

$$h_{sm}(n) = |h_{12}(n)| * h_{w1}(n) \tag{23}$$

  where $*$ denotes convolution operation and $h_{w1}(n)$ is Hamming window of 50 ms duration.
- Nonlinearly map smoothed error function values by using a sigmoid nonlinear function with slope parameter $\lambda = 20$ and T equal to 0·2 times average value of the normalized sum.
- The nonlinearly mapped values are termed as gross weight function.

Fine Level Processing

- Smooth the smaller variations in the error function using the Hamming window. That is,

$$h_s(n) = h_{12}(n) * h_{w2}(n) \tag{24}$$

  where $h_{w2}(n)$ is the Hamming window of 3 ms duration.
- The major peaks in the smoothed difference values $h_s(n)$ indicate approximate locations of instants of significant excitation. In particular, positive peaks correspond to the instants of desired speaker and negative peaks correspond to the instants of undesired speaker
- Convolve the negative of FOGD operator with the positive values of mean smoothed HE of the LP residual to determine the desired speaker instants location.
- Convolve the negative of FOGD operator with the negative values of mean smoothed HE of the LP residual to determine the interfering speaker instants location.
- Compute the fine weight function as

$$W_f(n) = [w_{\min} + (1 + w_{\min})W_a(n)] - w_{\min} W_b(n) \tag{25}$$

$$W_a(n) = \sum_{i=1}^{N_a} \delta(n - a_i) * h_{w2}(n) \tag{26}$$

Table 3. (Continued).

$$W_b(n) = \sum_{i=1}^{N_b} \delta(n - b_i) * h_{w2}(n) \tag{27}$$

where a_i and b_i represent the approximate locations of instants of significant excitation of desired and undesired speaker, respectively. N_a and N_b represent total number of detected instants of desired and undesired speaker, respectively. $w_{\min}$ is set as 0·3 and $h_{w2}(n)$ is the Hamming window of 3 ms duration.

Final Weight Function

- Multiply the two weight functions (gross and fine weight functions) to generate the final weight function.
- Multiply the LP residual signal of noisy speech by the final weight function.
- Excite the time-varying all-pole filter (derived from degraded speech) using weighted residual to obtain temporally processed speech.

Spectral Processing

- Compute the HE of enhanced LP residual (i.e. LP residual weighted by the weight function).
- Perform the autocorrelation on the HE of LP residual using a frame size of 40 ms and a frame shift of 10 ms. The normalized autocorrelation is obtained as (Proakis & Manolakis 1996)

$$R(l) = \frac{\sum_{n=0}^{L-1-l} h_m(n)h_m(n+l)}{\sum_{n=0}^{L-1} h_m^2(n)}; \quad l = 0, 1, 2, \ldots, L-1 \tag{28}$$

where $L = 320$ for $F_s = 8$ kHz and

$$h_m(n) = h(n) - E\{h(n)\} \tag{29}$$

where $E[.]$ is the expected value operator.
- Find the pitch estimate from the first major after the center peak in the range of 2·5 ms to 12·5 ms.
- Compute the similarity measure (S_m) and the magnitude of first major peak (R_p) in the normalized autocorrelation sequence.
- The similarity is measured by comparing samples in a region of 2 ms on either side of the first major peak of present frame with samples from previous/next frame. The similarity measure is computed as

$$S_m = \max\left\{\frac{COV(R_i, R_{i-1})}{\sigma_{R_i}\sigma_{R_{i-1}}}, \frac{COV(R_i, R_{i+1})}{\sigma_{R_i}\sigma_{R_{i+1}}}\right\} \tag{30}$$

where $COV(X, Y) = E(XY) - E(X)E(Y)$ and $\sigma_X = \sqrt{E(X) - E^2(X)}$. R_i, R_{i-1} and R_{i+1} represent samples around the first major peak in the current, previous and next frame, respectively.
- A frame of speech subjected to autocorrelation is considered as voiced frame only when the values of $R_p \geq 0{\cdot}4$ (Markel 1972) and $S_m \geq 0{\cdot}7$ (Prasanna & Yegnanarayana 2004).
- For voiced regions, sample and enhance the pitch and harmonics of the desired speaker.

$$X_s(k) = A_f \times X(k) \times W(k). \tag{31}$$

Table 3. (Continued).

where the multiplication factor A_f is used to further enhance the pitch and harmonics of the desired speaker with reference to undesired speaker (Harmonic sampling with spectral enhancement of desired speaker (Krishnamoorthy & Prasanna 2007)). In the present study A_f is chosen as 2 and

$$W(k) = P(k) * h_r(k) \tag{32}$$

$$P(k) = \sum_{i=1}^{N_p} \delta(k - p_i) \tag{33}$$

where P_i is the frequency indexes of harmonics and N_p represents total number of harmonics and

$$h_r(k) = \begin{cases} 1, & -2 \le k \le 2 \\ 0, & \text{otherwise} \end{cases} \tag{34}$$

For non-speech or unvoiced region

$$X_s(k) = X(k). \tag{35}$$

Reconstruct the enhanced speech signal by IDFT and OLA method.

calculation of MFCCs, the short term Fourier transform (STFT) analysis is performed on the speech signal using frame size of 20 ms with shift of 10 ms. For each frame the STFT magnitude spectrum is computed and is further processed by the 24 triangular shaped mel-filter banks to find out the filter bank energies. Then discrete cosine transform (DCT) is taken on the spectral energies to obtain MFCCs (Picone 1993). We have used a 13-dimensional MFCC vector (excluding c_0) appended with delta (Δ) and delta–delta ($\Delta\Delta$) coefficients as feature vector of each frame. The Δ and $\Delta\Delta$ coefficients obtained respectively from the MFCC and Δ by the first-order time derivative to capture the temporal dynamics of the signal (Furui 1981). The commonly used definition for computing Δ parameter is (Furui 1981)

$$\Delta c_m(n) = \frac{\sum_{i=-T}^{T} k_i c_m(n+i)}{\sum_{i=-T}^{T} |i|} \tag{36}$$

where $c_m(n)$ denotes the m^{th} feature for the n^{th} time frame, k_i is the i^{th} weight and T is the number of successive frames used for computation. Generally, T is taken as 2 (Furui 1981).

3.3 *Speaker modelling*

The Gaussian mixture model (GMM) employing universal background model (UBM) with MAP speaker adaptation is the dominant approach in text-independent speaker recognition (Reynolds 2000; Prakash & Hansen 2007). GMM basically models the feature vectors of the speaker as Gaussian densities. For a D-dimensional feature vector denoted as x_t, the mixture density for speaker Ω is defined as weighted sum of M component Gaussian densities as given by (Reynolds 1995)

$$P(x_t|\Omega) = \sum_{i=1}^{M} w_i P_i(x_t), \tag{37}$$

where w_i are the weights and $P_i(x_t)$ are the component densities. Each component density is a D-variate Gaussian function of the form

$$P_i(x_t) = \frac{1}{(2\pi)^{D/2}\,|\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}[(x_t-\mu_i)'\Sigma_i^{-1}(x_t-\mu_i)]} \tag{38}$$

where μ_i is a mean vector and Σ_i covariance matrix for i^{th} component. The mixture weights have to satisfy the constraint $\sum_{i=1}^{M} w_i = 1$. The complete Gaussian mixture density is parameterized by the mean vector, the covariance matrix and the mixture weight from all component densities.

Before using GMM for speaker identification, the model parameters must be estimated. There are many criteria that can be used to estimate the model parameters. The most common one is the maximum likelihood (ML) criterion. Of the many techniques developed to maximize the likelihood value, the most popular is the iterative expectation maximization (EM) algorithm. Generally, the log-likelihood is preferred since it is computationally faster to process (Bimbot *et al* 2004). For a sequence of training vectors $X = \{x_1, x_2, \ldots, x_T\}$, under the assumption of independent feature vectors, the log-likelihood function can be written as (Reynolds 1995)

$$\log[P(X|\Omega)] = \sum_{t=1}^{T} \log\left[\sum_{i=1}^{M} w_i P_i(x_t)\right]. \tag{39}$$

However, it is necessary that sufficient training data is available in order to create a model of the speaker. Another way of estimating a statistical model, which is especially useful when the training data available is of short duration, is by using maximum *a posteriori* (MAP) adaptation of the universal background model (UBM) trained on the speech data of several other speakers (Reynolds 2000). This UBM is trained with large amount of data which covers the different kinds of speech that may be encountered by the system during training.

The UBM–GMM can be mainly divided into three parts: UBM training, Bayesian adaptation of speaker models and speaker identification. A speaker-independent model, or UBM, is trained from the non-target speakers using the EM algorithm. Then for each target speaker, speaker models are created through MAP adaptation of the UBM using speaker-specific training speech. Based on experimental results, the best performance can be achieved using only the mean adaptation of Gaussian mixtures (Reynolds 2000; Bimbot *et al* 2004; Prakash & Hansen 2007). The mean adaptation equations are given by (Reynolds 2000)

$$\hat{\mu}_i = \frac{n_i}{n_i + r} E_i(x_t) + \frac{r_i}{n_i + r_i} \mu_i, \tag{40}$$

where μ_i mean value of UBM and

$$n_i = \sum_{i=1}^{T} \mathrm{P}(i|x_t) \tag{41}$$

$$E_i(x_t) = \frac{\sum_{i=1}^{T} \mathrm{P}(i|x_t)x_t}{n_i}, \tag{42}$$

where r is a fixed relevance factor that controls the balance of adaptation between the UBM parameters and speaker-specific training observations. It is set to 16 as in (Reynolds 2000).

3.4 *Testing*

In identification phase, mixture densities are calculated for every feature vector for all speakers and speaker with ML is selected as identified speaker (Reynolds 1995). For example, if S speaker models $\{\Omega_{1,}\Omega_{2,\ldots,}\Omega_S\}$ are available after the training, the sequence of test feature vectors $X = \{x_1, x_2, \ldots, x_T\}$ is tested by finding the speaker model which has maximum *a posteriori* probability. According to the Bayes rule (Reynolds 1995)

$$\hat{s} = \max_{1\le s\le S} P(\Omega_S|X) = \max_{1\le s\le S} \frac{P(X|\Omega_S)}{P(X)} P(\Omega_S). \tag{43}$$

Assuming equal probability of all speakers and the statistical independence of the observations, the decision rule for the most probable speaker can be redefined as (Bimbot *et al* 2004)

$$\hat{s} = \max_{1\le s\le S} \sum_{t=1}^{T} \log P(x_t|\Omega_s) \tag{44}$$

where T is the number of feature vectors of the speech data set under test. Finally, the recognition accuracy of speaker identification is assessed by identification rate and is defined as the ratio of the number of correctly identified utterances to the total number of testing utterances.

4. Experimental results and discussions

In this work, first the speaker models are created using clean speech data by UBM–GMM concept. The UBM is trained on approximately one hour of data (excluding silence regions) of the TIMIT database. The UBM training data are taken from *train* set of the TIMIT database and for evaluation speech samples are taken from the *test* set of the TIMIT database. A first set of experiment is conducted to evaluate the performance of the speaker recognition system under clean condition and is found to be 97%. In the next step to evaluate the performance of the combined TSP method on degraded speech the approach followed is schematically illustrated in figure 3. In figure the pre-processing block refers to temporal processing, spectral processing and the combined TSP methods. The recognition studies are carried out on test speech (degraded speech) data with and without pre-processing.

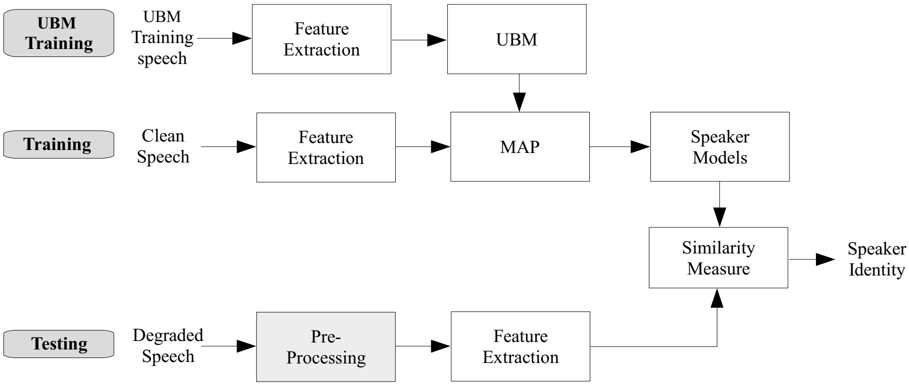

Figure 3. Block diagram of speaker recognition under degraded conditions.

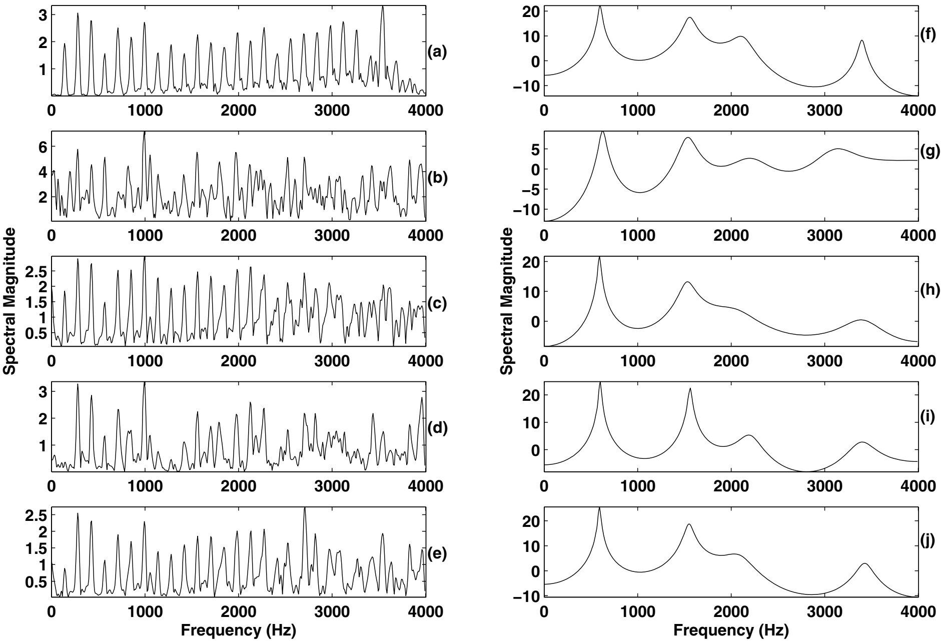

Figure 4. Noisy speech enhancement: Excitation source spectrum of a frame of **(a)** clean speech, **(b)** degraded speech, **(c)** speech processed by temporal processing, **(d)** speech processed by spectral processing, **(e)** speech processed by temporal and spectral processing, and Vocal tract (LP) spectrum of a frame of **(f)** clean speech, **(g)** degraded speech, **(h)** speech processed by temporal processing, **(i)** speech processed by spectral processing, and **(j)** speech processed by temporal and spectral processing.

4.1 *Speaker recognition in noisy environment*

The noisy speech is created by adding white noise from NOISEX-92 database (Varga & Steeneken 1993) to the test utterances. The noise waveform being added to the speech was scaled to give the desired global SNR. The value of SNR is varied over the range of 0–30 dB. While testing the degraded speech signal is enhanced using individual and the proposed TSP methods as given in table 1 prior to the feature extraction step. In the present work, the conventional multi-band spectral subtraction (Kamath & Loizou 2002) and the MMSE–STSA estimator (Ephraim & Malah 1984) methods are used for combination. For MMSE–STSA estimator the *a priori* SNR for each frequency component is estimated using a decision directed variance estimator approach (Ephraim & Malah 1984). In this approach, the variance estimator at a given frame uses the signal spectral magnitude estimate from the previous frame along with the current noisy spectral values.

To illustrate the merit of the combined TSP, figures 4a–e show the excitation source signal (LP residual) spectrum of clean, degraded, temporal, spectral and combined temporal and spectral processed speech, respectively. Similarly, figures 4f–j show the vocal spectrum of clean, degraded, temporal, spectral and combined temporal and spectral processed speech, respectively. The combined processing shows improvement in both the excitation source and vocal tract spectrum, where as individual processing methods show major improvement either at the excitation source signal or at the vocal tract spectrum only. As a result the extracted speaker-specific features of combined TSP method are more robust compared to the individual processing methods and thus result in improved speaker recognition performance.

Table 4 shows experimental results obtained for the various SNR level. It can be seen that the recognition performance of the combined TSP is higher than individual processing methods. However, for higher SNR values, in particular for 30 dB, the combined method results slightly lower performance than spectral processing alone. This is mainly because the underlying temporal processing method involves weighting of the excitation source signal for the enhancement. For higher SNR values, since noise level is very low, the weighting may disturb the actual signal and thus results in the slight reduction in performance.

4.2 *Speaker recognition in reverberant environment*

The reverberant speech signal is generated through a linear convolution between the original speech data and a room impulse response. We have generated five impulse responses using image method (Allen & Berkley 1979) having the reverberation time in the range of 0·2–1·0 sec with a source microphone distance of 1·5 m. The same procedure which is followed in the noisy speech experiment is repeated and the obtained results are given in table 5. The combined TSP algorithm used for enhancing the reverberant speech is given in table 2. It can be observed that the combined TSP method gives relatively higher performance than individual processing method. This is mainly because (Krishnamoorthy & Prasanna 2009)

(i) Spectral processing removes the late reverberant portion of the reverberant speech.
(ii) Temporal processing enhances the early reverberant portion of the reverberant speech.

Figures 5a and b show the clean speech and the corresponding reverberant speech obtained by convolving the impulse response from the image method (Allen & Berkley 1979) with a reverberation time of approximately 1 sec. The speech processed by the temporal processing, spectral processing and the proposed combined TSP are given in figures 5c–e, respectively. The spectrograms of the respective signals shown in figures 5a–e are given in figures 5f–j. All the spectrograms are constructed using Hamming window of 128 samples with shift of 64 samples. From figure 5 it can be inferred that the spectral processing removes the late reverberant speech spectral components where as the temporal processing fails to remove the late reverberation in some portions. Figures 6c–e show the nature of the excitation source signal obtained from the different processing methods. A small segment of voiced portion (high SRR region) is taken for illustration. It can be noticed that, with reference to clean speech instant locations (pointed by the down arrow symbol in figure 6a), the excitation source signal of spectral processed speech does not show any noticeable improvement with reference to the degraded one. On the other hand, the temporal processing shows the improvement around the major instant locations in the excitation source information. It can be seen from figures 5e and 6e that the combined TSP method provides enhancement of high SRR regions and also the suppression of late reverberation. As a result, speech processed by the combined TSP method gives lesser spectral distance with reference to clean speech (Krishnamoorthy & Prasanna 2009) and therefore extracted MFCC features are more closer to that of the clean speech MFCC features. This leads to the improvement in the speaker identification rate.

4.3 *Speaker recognition in two-speaker environment*

For this study, first a set of 100 speakers different from the UBM training and testing set is chosen as interfering speakers. The synthetic two speaker data for a pair of microphones are created with delays of $d_1 = 8$ and $d_2 = -16$ samples and used as a test signal for speaker recognition study. Two different types of two speaker data are created, they are; (i) the gender

Table 4. Speaker recognition performance (%) under noisy environment. In table abbreviations DEG, TP, SP1, SP2, TSP1 and TSP2 refer to degraded speech, temporal processing, multi band spectral subtraction (Kamath & Loizou 2002), MMSE–STSA estimator (Ephraim & Malah 1984), combined temporal and multi-band spectral subtraction and combined temporal and MMSE–STSA estimator, respectively. P_i represents the maximum performance among the number of Gaussians.

	No. of Gaussians									No. of Gaussians							
	8	16	32	64	128	256	512	Max (P_i)		8	16	32	64	128	256	512	Max (P_i)
Clean	85·0	92·5	94·0	95·5	95·5	97·0	96·5	97·0	Clean	85·0	92·5	94·0	95·5	95·5	97·0	96·5	97·0
SNR = 0 dB									SNR = 12 dB								
DEG	0·5	1·0	0·5	0·5	1·5	1·5	1·5	1·5	DEG	5·5	5·0	6·0	8·0	8·5	10·5	9·5	10·5
TP	1·0	1·0	1·0	1·5	2·5	3·0	2·5	3·0	TP	18·5	24·0	29·0	32·5	33·5	33·0	30·0	33·5
SP1	4·5	5·0	6·0	3·5	5·0	6·0	6·0	6·0	SP1	38·5	39·5	49·5	44·5	45·0	43·0	40·5	49·5
SP2	4·5	4·5	3·5	2·5	5·5	4·0	5·5	5·5	SP2	43·0	48·0	55·5	53·0	54·5	50·0	57·0	57·0
TSP1	8·5	8·0	10·5	12·5	10·5	9·5	13·0	13·0	TSP1	54·0	56·5	67·0	66·5	71·0	72·0	72·5	72·5
TSP2	4·0	6·0	7·5	8·0	7·5	6·5	7·5	8·0	TSP2	55·0	56·5	66·0	67·0	66·5	63·5	67·5	67·5
SNR = 3 dB									SNR = 15 dB								
DEG	1·0	1·0	1·0	1·5	2·0	2·0	2·0	2·0	DEG	13·0	12·0	17·0	20·0	19·5	22·0	23·5	23·5
TP	1·0	1·0	2·0	1·5	3·5	5·0	2·5	5·0	TP	27·0	37·0	41·5	42·5	42·5	40·5	40·5	42·5
SP1	7·5	12·0	11·0	13·5	14·5	15·0	19·0	19·0	SP1	56·5	60·5	65·0	67·0	69·5	70·5	68·0	70·5
SP2	5·5	8·0	6·5	11·0	10·0	8·0	12·0	12·0	SP2	60·0	67·5	70·5	77·0	77·0	75·5	74·0	77·0
TSP1	13·0	18·0	22·5	17·5	22·0	20·5	24·0	24·0	TSP1	60·0	65·5	76·5	77·5	83·0	82·5	79·0	83·0
TSP2	8·5	13·5	13·5	12·0	11·0	9·0	11·0	13·5	TSP2	61·0	67·5	73·0	77·0	77·0	80·5	78·5	80·5

Table 4. (Continued).

	No. of Gaussians									No. of Gaussians							
	8	16	32	64	128	256	512	Max (P_i)		8	16	32	64	128	256	512	Max (P_i)
				SNR = 6 dB									SNR = 20 dB				
DEG	1·0	1·0	2·0	2·0	2·0	2·0	2·0	2·0	DEG	26·5	35·0	38·5	49·0	50·0	51·5	51·0	51·5
TP	2·5	7·0	6·0	13·5	10·5	16·0	12·5	16·0	TP	52·0	60·5	71·0	72·5	78·5	76·0	76·0	78·5
SP1	16·5	21·5	27·0	27·0	32·0	28·0	31·5	32·0	SP1	73·5	79·5	82·5	85·5	87·0	87·0	86·0	87·0
SP2	9·0	15·0	16·5	20·0	19·0	15·0	21·0	21·0	SP2	71·0	76·0	80·0	86·5	84·0	86·0	85·0	86·5
TSP1	23·0	29·0	32·0	31·0	36·5	38·0	41·5	41·5	TSP1	67·0	75·5	78·5	83·5	88·0	86·5	84·5	88·0
TSP2	19·5	24·5	27·5	30·0	28·5	24·0	27·5	30·0	TSP2	70·5	76·0	80·5	85·5	85·5	87·0	83·5	87·0
				SNR = 9 dB									SNR = 30 dB				
DEG	3·0	2·5	3·0	3·5	3·0	2·5	2·0	3·5	DEG	59·0	72·5	79·0	86·5	87·0	86·0	89·0	89·0
TP	11·0	13·0	15·5	17·0	18·5	21·0	19·5	21·0	TP	66·0	76·0	82·0	86·0	86·5	87·5	85·0	87·5
SP1	31·0	39·0	44·5	42·0	49·5	48·5	49·0	49·5	SP1	80·0	84·0	86·0	92·0	91·0	90·0	90·5	92·0
SP2	22·5	25·0	32·5	32·0	33·5	28·0	36·0	36·0	SP2	78·0	82·5	85·0	91·5	90·5	91·0	91·5	91·5
TSP1	36·0	43·0	46·0	54·5	59·0	57·0	59·0	59·0	TSP1	71·5	78·5	83·0	86·0	87·5	88·5	89·0	89·0
TSP2	37·0	40·0	48·5	50·5	48·0	52·5	53·5	53·5	TSP2	71·5	79·0	82·5	89·0	87·0	88·0	87·0	89·0

Table 5. Speaker recognition performance (%) under reverberant environment. In the table abbreviations DEG, TP, SP and TSP refer to degraded speech, temporal processing, spectral processing and combined temporal and spectral processing, respectively. P_i represents the maximum performance among the number of Gaussians. D and T_{60} represent source microphone distance and reverberation time, respectively.

	No. of Gaussians							
	8	16	32	64	128	256	512	Max (P_i)
D = 1·5 m and T_{60} = 0·2 sec								
DEG	29·0	32·0	36·0	35·0	43·0	38·0	37·0	43·0
SP	42·0	45·5	51·5	56·5	53·0	54·0	53·5	56·5
TP	35·0	34·0	37·0	35·5	43·0	36·0	37·5	43·0
TSP	38·5	44·5	49·5	51·5	52·0	56·5	52·5	56·5
D = 1·5 m and T_{60} = 0·4 sec								
DEG	24·5	23·0	28·5	31·0	26·5	30·5	30·5	31·0
SP	36·0	35·0	40·0	45·5	45·0	49·5	47·0	49·5
TP	26·0	24·5	30·0	32·0	32·5	34·5	33·5	34·5
TSP	37·0	37·0	39·0	44·0	46·0	53·5	46·5	53·5
D = 1·5 m and T_{60} = 0·6 sec								
DEG	18·5	20·5	21·5	20·0	16·0	23·0	22·5	23·0
SP	17·0	18·0	23·0	21·5	17·0	26·0	21·5	26·0
TP	23·0	26·5	27·0	32·5	30·0	35·0	29·0	35·0
TSP	23·0	26·0	30·5	34·0	32·5	39·5	31·5	39·5
D = 1·5 m and T_{60} = 0·8 sec								
DEG	13·5	15·5	16·0	15·5	18·0	18·5	19·0	19·0
SP	20·0	18·0	18·5	23·0	23·0	24·5	24·5	24·5
TP	16·5	15·0	21·0	17·5	18·5	19·5	21·0	21·0
TSP	21·0	19·5	19·0	24·0	31·5	29·5	25·0	31·5
D = 1·5 m and T_{60} = 1·0 sec								
DEG	8·0	5·5	7·5	8·5	8·0	9·0	7·5	9·0
SP	12·5	14·0	17·5	19·0	14·5	19·0	19·0	19·0
TP	8·0	7·5	7·5	8·0	9·5	9·0	8·5	9·5
TSP	17·5	15·0	21·0	20·0	15·0	21·5	18·0	21·5

of the interfering speaker is same as that of original test speaker, and (ii) interfering speaker gender is different from that of test speaker. The degraded signal is pre-processed according to the method given in table 3 and the pre-processed speech signal is used for capturing MFCC features. Table 6 shows the result of this study for different pre-processing techniques. The relative improvement in the performance of speaker recognition with the different pre-processing techniques is clearly seen from the results. It can be observed that identification rate of same gender case is less than the different gender case. This may be interpreted in following way. In all pitch based separation methods, speech separation not only depends on the processing method used but also on the nature of the degraded signal. The more separated in the pitch and harmonics of each talker, the better the result to be expected. For same gender case the separation between the pitch of desired and interfering speaker may be minimum. Therefore for this case, the temporal and spectral processing may leave some of the interfering speaker information in the enhanced speech and thus results a slightly poor performance than different gender case.

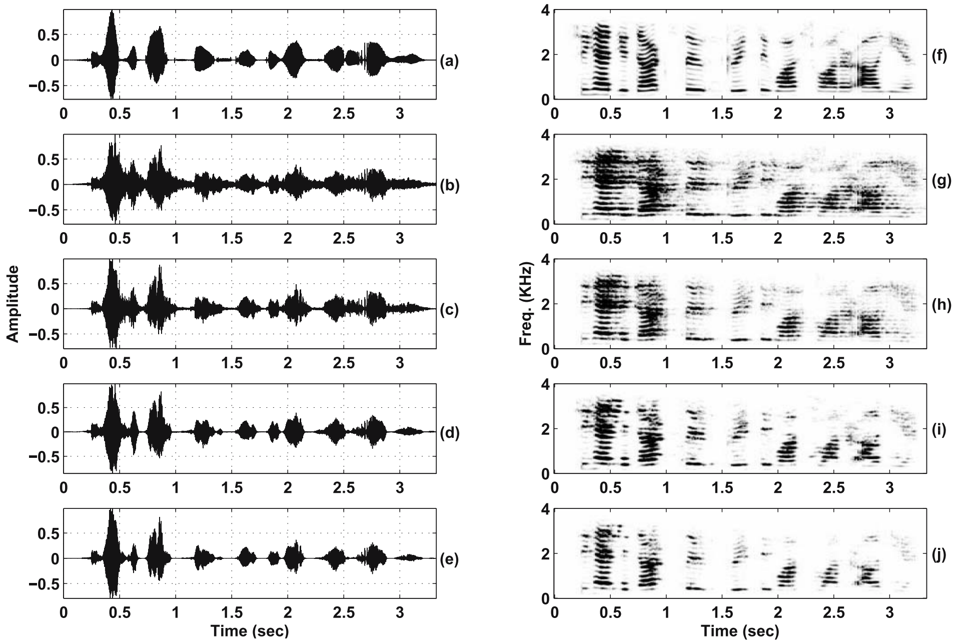

Figure 5. Reverberant speech enhancement: **(a)** clean speech, **(b)** degraded speech, **(c)** speech processed by temporal processing, **(d)** speech processed by spectral processing, **(e)** speech processed by spectral and temporal processing and **(f)–(j)** spectrograms of the respective signals shown in **(a)–(e)**.

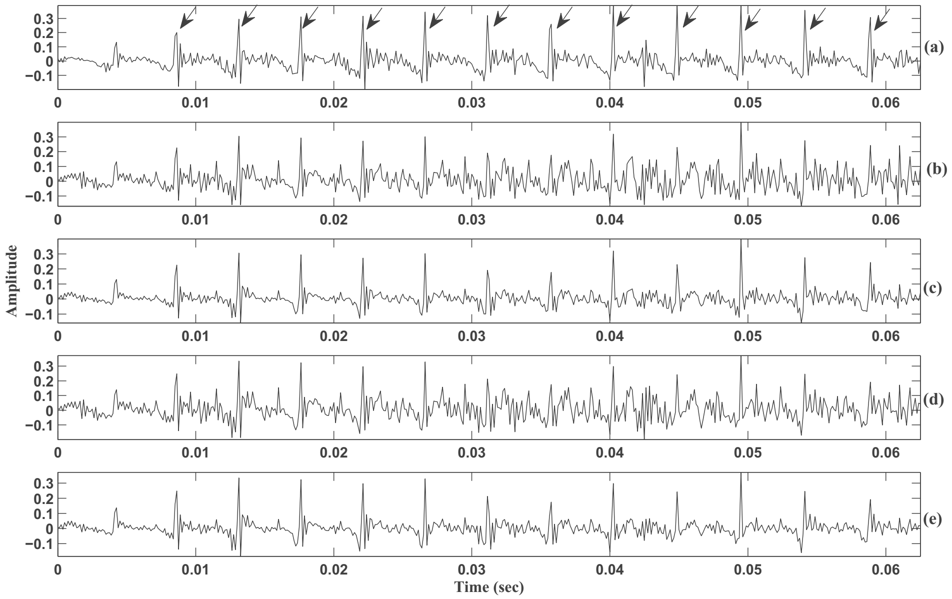

Figure 6. Reverberant speech enhancement (Krishnamoorthy & Prasanna 2009): Excitation source signal of **(a)** clean speech, **(b)** reverberant speech, **(c)** speech processed by temporal processing, **(d)** speech processed by spectral processing, and **(e)** speech processed by temporal and spectral processing.

Table 6. Speaker recognition performance (%) in two speaker environment. In the table abbreviations DEG, TP, SP and TSP refer to degraded speech, temporal processing, spectral processing and combined temporal and spectral processing, respectively. P_i represents the maximum performance among the number of Gaussians.

	No. of Gaussians							
	8	16	32	64	128	256	512	Max (P_i)
				Same Gender				
DEG	18·5	22·5	29·0	32·0	37·0	36·0	39·5	39·5
TP	26·0	35·0	38·0	44·0	45·0	50·5	50·0	50·5
SP	27·5	36·5	41·5	50·0	56·0	51·5	56·0	56·0
TSP	31·0	45·0	45·5	49·5	59·0	52·0	55·0	59·0
				Different Gender				
DEG	28·0	31·0	38·0	38·0	41·5	40·0	40·0	41·5
TP	32·0	40·0	43·0	45·0	52·0	50·0	49·0	52·0
SP	41·0	47·5	48·5	57·0	55·5	58·0	55·0	58·0
TSP	42·0	49·5	50·0	58·5	59·5	62·5	60·0	62·5

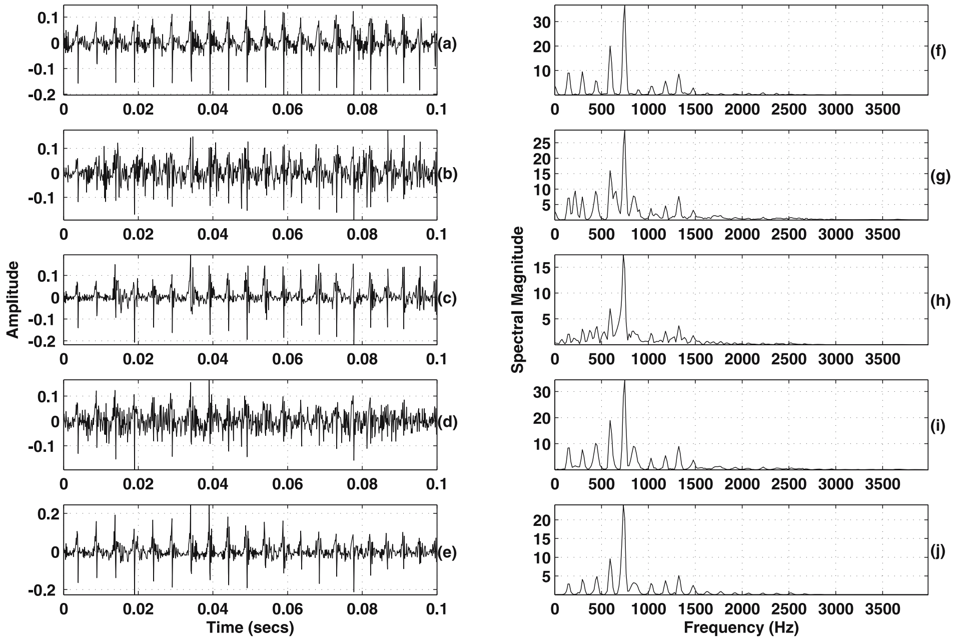

Figure 7. Two speaker speech separation: Excitation source signal of **(a)** clean speech of desired speaker, **(b)** degraded speech, **(c)** speech processed by temporal processing, **(d)** speech processed by spectral processing, **(e)** speech processed by temporal and spectral processing, and STFT magnitude spectrum of **(f)** clean speech of desired speaker, **(g)** degraded speech, **(h)** speech processed by temporal processing, **(i)** speech processed by spectral processing, and **(j)** speech processed by temporal and spectral processing.

To show the effectiveness of combined method in more clear way, figures 7a–e show the excitation source signal of desired speaker, degraded, temporal, spectral and combined temporal and spectral processed speech, respectively. Similarly, figures 7f–j show the STFT magnitude spectrum of clean, degraded, temporal, spectral and combined temporal and spectral processed speech, respectively. Here also it can be observed that the combined processing shows improvement in both the excitation source signal and short time spectrum, where as individual processing methods show major improvement either at the excitation source signal or at the short term spectrum only. Hence the combined TSP method gives better performance than individual processing method.

5. Summary and conclusions

The main objective of this work is to evaluate the performance of the combined TSP based speech enhancement methods in the speaker recognition task. For this study the TIMIT database is taken and MFCC features are extracted from the clean speech data and the speaker models are trained using UBM–GMM system. In testing stage, the synthetic degraded speech is generated for each type of degradation. The degraded speech is subjected to temporal, spectral and combined TSP methods before the feature extraction step to attenuate the degradation characteristics. The enhanced speech is subjected to testing. The recognition results show that combined processing method gives relatively higher performance than individual processing methods, except with some limitations. Like in very high SNR values and for lower reverberation times the combined method results slightly lower or equal performance than expected due to underlying processing steps involved in temporal processing.

Appendix A

Enhancement of high SNR Evidence

The evidence about the high SNR regions is enhanced by computing its slope with the help of a first order difference (FOD). The steps involved in the enhancement of high SNR evidences are explained for the parameter sum of peaks in the DFT spectrum with the help of figure 8. For illustration, speech data spoken by a female speaker and sampled at 8 kHz with a resolution of 16 bits/sample is taken and white Gaussian noise is added to make SNR of the signal is 5 dB. Here, SNR is defined as the ratio of the total power of the clean speech to the total power of the additive background noise. The sum of largest ten peaks of the DFT spectrum of the Hamming windowed signal is calculated using window of 20 ms duration and 10 ms overlap between the frames. All these values are repeated for frame shift number of times to make the sequence length equal to that of the speech signal (Krishnamoorthy & Prasanna 2009). Figure 8b shows the normalized sum of peaks in the DFT spectrum for the noisy signal given in figure 8a. Since FOD represents the slope, the positive to negative going zero transition in FOD locates the peaks in the sum of DFT spectrum values. The positive to negative going zero transition points and the corresponding local peaks are represented by star (*) symbols in figures 8b and c. The unwanted zero crossings that are detected at the low SNR regions are eliminated by finding the sum of absolute FOD values for a duration of 5 ms on either side with reference to each positive to negative going zero crossing point and are given in figure 8d. The peaks with the lower FOD values are eliminated by setting the threshold at 0·5 times the mean values of the FOD. In the next step, if two successive peaks occur within

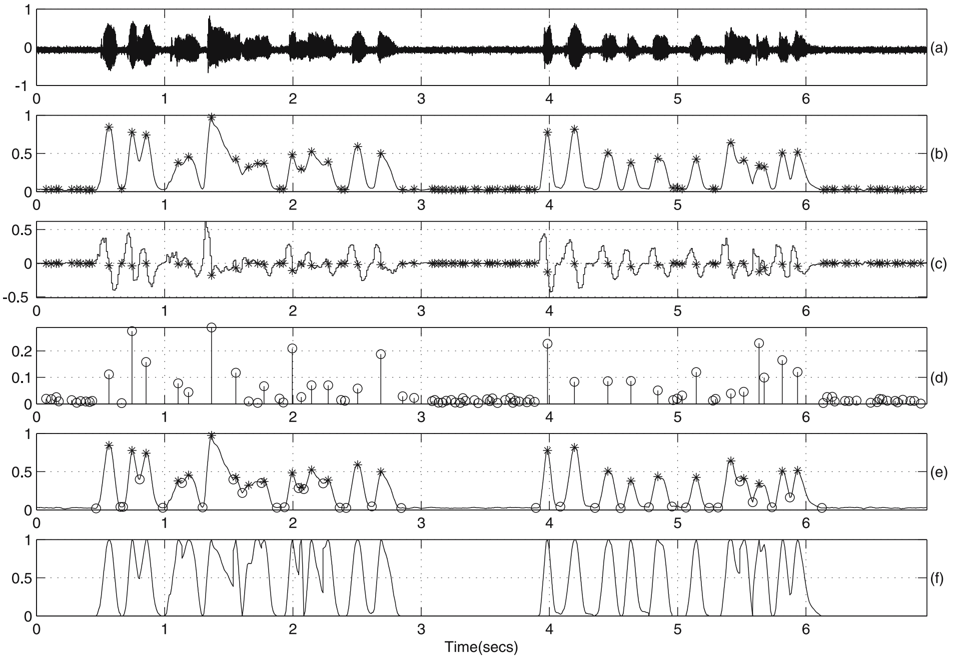

Figure 8. High SNR regions enhancement: **(a)** noisy speech, **(b)** normalized sum of peaks in the DFT spectrum, **(c)** first order difference (FOD) values, **(d)** sum of absolute FOD values computed for a duration of 5 ms on either side with reference to each positive to negative going zero crossing point, **(e)** sum of peaks in the DFT spectrum and high SNR region locations and **(f)** enhanced sum of peaks in the DFT spectrum values. In the figures, * and o represent the peaks and their boundaries of high SNR regions, respectively.

50 ms then the peak with the lower FOD value is eliminated based on the assumption that occurrence of two high SRR regions unlikely within a 50 ms interval. The star (*) symbols in figure 8e show the peak locations after eliminating the undesirable peaks (Krishnamoorthy & Prasanna 2009). Segments bounded by negative to positive going zero transition points on either side are enhanced by normalizing so that the peak smoothed DFT in each such region has an amplitude of 1·0, as shown in figure 8f. As it can be observed, the change at the SNR is further enhanced by processing sum of peaks in the DFT spectrum.

References

Allen J, Berkley D 1979 Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am.* 65: 943–950

Bimbot F, Bonastre J F, Fredouille C, Gravier G, Chagnolleau M I, Meignier S, Merlin T, Garcia O J, Delacretaz P, Reynolds 2004 A tutorial on text-independent speaker verification. *EURASIP J. Applied Signal process.* 4: 430–451

Boll S 1979 Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust., Speech, Signal process.* ASSP-27 113–120

Campbell J P 1997 Speaker recognition: A tutorial. *Proc. IEEE* 85(9): 1437–1462

Ephraim Y, Malah D 1984 Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal process.* ASSP-32 1109–1121

Furui S 1981 Comparison of speaker recognition methods using statistical features and dynamic features. *IEEE Trans. Audio, Speech and Language process.* 29(3): 342–350

Greenberg S, Kingsbury B E D 1997 The modulation spectrogram: in pursuit of an invariant representation of speech. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal process.* Munich, Germany 1647–1650

Habets E A P, Gannot S, Cohen I, Sommen P C W 2008 Joint dereverberation and residual echo suppression of speech signals in noisy environments. *IEEE Trans. Audio, Speech, and Language Process.* 16(8): 1433–1451

Heck L P, Konig Y, Sönmez M K, Weintraub M 2000 Robustness to telephone handset distortion in speaker recognition by discriminative feature design. *Speech Communication* 31(2–3): 181–192

Kamath S, Loizou P 2002 A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal process.* Orlando, USA

Krishnamoorthy P, Prasanna S R M 2007 Processing noisy speech for enhancement. *J. IETE Technical Review*, Special issue on spoken language processing 24: 349–355

Krishnamoorthy P, Prasanna S R M 2008 Temporal and spectral processing of degraded speech. In: *IEEE Proc. Int. Conf. Advanced Computing and Communications* 112–118

Krishnamoorthy P, Prasanna S R M 2009 Reverberant speech enhancement by temporal and spectral processing. *IEEE Trans. Speech, Audio and Language Process.* 17(2): 253–266

Lebart K, Boucher J 2001 A new method based on spectral subtraction for speech dereverberation. *Acta Acoustica* 87: 359–366

Markel J 1972 The SIFT algorithm for fundamental frequency estimation. *IEEE Trans. Audio and Electroacoustics* 20: 367–377

McAulay R, Quatieri T 1986 Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoust., Speech, Signal process.* ASSP-34 744–754

Ming J, Hazen T, Glass J, Reynolds D 2007 Robust speaker recognition in noisy conditions. *IEEE Trans. Audio, Speech, and Language process.* 15(5): 1711–1723

Morgan D, George E, Lee L, Kay S 1997 Cochannel speaker separation by harmonic enhancement and suppression. *IEEE Trans. Speech Audio process.* 5: 407–424

Murty K S R, Yegnanarayana B 2008 Epoch extraction from speech signals. *IEEE Trans. Audio, Speech, and Language Process.* 16(8): 1602–1613

Ortega-Garcia J, Gonzalez-Rodriguez J 1996 Overview of speech enhancement techniques for automatic speaker recognition. In: *Proc. Fourth Int. Conf. Spoken Language.* 2: 929–932

Parsons T 1976 Separation of speech from interfering speech by means of harmonic selection. *J. Acoust. Soc. Am.* 60: 911–918

Picone J 1993 Signal modelling techniques in speech recognition. *Proc. IEEE* 81(9): 1215–1247

Prakash V, Hansen J 2007 In-set/out-of-set speaker recognition under sparse enrollment. *IEEE Trans. Audio, Speech, and Language process.* 15(7): 2044–2052

Prasanna S R M, Sandeep Reddy B, Krishnamoorthy P 2009 Vowel onset point detection using source, spectral peaks and modulation spectrum energies. *IEEE Trans. Speech, Audio and Language Process.* 17(4): 556–565

Prasanna S R M, Subramanian A 2005 Finding pitch markers using first order Gaussian differentiator. In: *IEEE Proc. Third Int. Conf. Intelligent Sensing Information Process.* 140–145

Prasanna S R M, Yegnanarayana B 2004 Extraction of pitch in adverse conditions. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal process.* Vol. 1. Montreal, Quebec, Canada I–109–I–112

Proakis J G, Manolakis D G 1996 Digital signal processing-principles, algorithms, and applications, 3rd Edition. Prentice Hall

Rao K, Prasanna S R M, Yegnanarayana B 2007 Determination of instants of significant excitation in speech using Hilbert envelope and group delay function. *IEEE Signal process. Letters* 14(10): 762–765

Reynolds D 1994 Experimental evaluation of features for robust speaker identification. *IEEE Trans., Speech Audio process.* 2(4): 639–643

Reynolds D A 1995 Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication* 17: 91–108

Reynolds D A 2000 Speaker verification using adapted Gaussian mixture models. *Digital Signal Process.* 10(1–3): 19–41

Sankar A, Lee C-H 1996 A maximum-likelihood approach to stochastic matching for robust speech recognition. *IEEE Trans. Speech Audio process.* 4(3): 190–202

Varga A, Steeneken H J M 1993 Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effct of additive noise on speech recognition systems. *Speech Communication* 12(3): 247–251

Wu M, Wang D 2006 A two-stage algorithm for one-microphone reverberant speech enhancement. *IEEE Trans. Audio, Speech, Language process.* 14: 774–784

Yegnanarayana B, Avendano C, Hermansky H, Satyanarayana Murthy P 1999 Speech enhancement using linear prediction residual. *Speech Communication* 28: 25–42

Yegnanarayana B, Prasanna S R M, Duraiswami R, Zotkin D 2005 Processing of reverberant speech for time-delay estimation. *IEEE Trans. Speech Audio process.* 13: 1110–1118

Yegnanarayana B, Prasanna S R M, Mathew M 2003 Enhancement of speech in multispeaker environment. In: *Proc. European Conf. Speech process., Technology.* Geneva, Switzerland 581–584

Yegnanarayana B, Satyanarayana Murthy P 2000 Enhancement of reverberant speech using LP residual signal. *IEEE Trans. Speech Audio process.* 8: 267–281

Zilovic M, Ramachandran R, Mammone R 1998 Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer functions. *IEEE Trans. Speech Audio process.* 6(3): 260–267

Zue V, Seneff S, Glass J 1990 Speech database development at MIT: TIMIT and beyond. *Speech Communication* 9(4): 351–356