Temporal and Spectral Processing Methods for Processing of Degraded Speech: A Review

P. Krishnamoorthy and S. R. Mahadeva Prasanna

Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, Guwahati-781 039, Assam, India.

Abstract

This paper presents an overview of several most commonly used temporal and spectral processing methods for the enhancement of degraded speech. Three major sources of degradation, namely, background noise, reverberation and speech from the competing speakers, have been considered. Temporal processing refers to processing the degraded speech in the time domain, for enhancing the speech components. Spectral processing refers to processing the degraded speech in the frequency domain. After the review, the paper concludes with a summation that considers the possibility of combined Temporal and Spectral Processing (TSP) approach for the enhancement of degraded speech.

Keywords

Multi-speaker speech, Noisy speech, Reverberant speech, Speech enhancement, Temporal processing and spectral processing.

1. Introduction

Speech signals in the real world scenarios are often corrupted by various types of degradations. The most common degradations include background noise, reverberation and speech from competing speaker(s). Degraded speech is poor, both in terms of perceptual quality and intelligibility. Poor perceptual quality leads to listener fatigue. Poor intelligibility leads to degraded performance in tasks like speech and speaker recognition. Degraded speech, therefore, needs to be processed for enhancing its perceptual quality and intelligibility. Several methods have been proposed in the literature for the enhancement of degraded speech. A majority of these methods can be grouped into *spectral processing* and *temporal processing* methods. In the spectral processing methods, degraded speech is processed in the frequency domain, for achieving enhancement. In the temporal processing methods, the processing is done in the time domain. The approach to speech enhancement varies considerably, depending on the type of degradation. For example, the type of processing for enhancing speech degraded by background noise (noisy speech) is different from the method employed for enhancing speech degraded by reverberation (reverberant speech).

This paper provides an overview of some commonly used methods proposed for the enhancement of degraded speech. The rest of the paper is organized as follows: Section 2 presents a review of the methods for processing speech degraded by background noise. Section 3 discusses the enhancement techniques for speech degraded by reverberation. Methods for the enhancement of speech in multi-speaker environment are discussed in Section 4. Finally, the summary of the review has been provided in Section 5.

2. Enhancement of Noisy Speech

Background noise is the most common factor that causes degradation of quality and intelligibility of speech. The term background noise refers to any unwanted signal that is added to the desired signal. Background noise can be stationary or nonstationary and is assumed to be uncorrelated and additive to the speech signal. Mathematically, speech degraded by background noise can be expressed as the sum of clean speech and background noise [1].

That is,

$$y(n) = s(n) + d(n) \quad (1)$$

where y(n), s(n) and d(n) denote the noisy speech, clean speech and background noise, respectively. In the frequency domain, it can be represented as

$$Y(k) = S(k) + D(k) \quad (2)$$

where k is the index of frequency bin.

The problem of enhancing noisy speech received considerable attention in the literature and a variety of methods have been proposed to overcome it. A majority of these methods may belong to one of these two categories: *Spectral processing methods* such as the spectral subtraction, minimum mean square error (MMSE) estimator, wavelet denoising methods and *temporal processing methods* such as linear prediction (LP) residual based methods.

2.1 Spectral Processing Methods

Spectral processing methods are the most popular techniques for noise reduction, mainly because of their simplicity and effectiveness. Most of the spectral processing techniques rely on the basis that the human speech perception is not sensitive to short-time phase [2]. This is exploited in these methods, where only the spectral magnitude associated with the original signal is estimated. In case of noisy speech, the spectral processing methods can be grouped into nonparametric and statistical model-based methods [2]. Methods from the first category remove an estimate of the degradation from the noisy features, such as subtractive type algorithms and wavelet denoising. The statistical model based method such as MMSE estimator uses the parametric model of the signal generation process.

2.1.1 Spectral Subtraction

Spectral subtraction is historically one of the first algorithms proposed in the field of background noise reduction, which is still referenced today because of its minimal complexity and relative ease in implementation. Spectral subtraction is performed by subtracting the average magnitude of the noise spectrum from the spectrum of the noisy speech, to estimate the magnitude of the enhanced speech spectrum [1]. The estimate of the enhanced speech spectrum is obtained as [1]:

$$\left|\hat{S}(k)\right| = \left|Y(k)\right| - \left|\hat{D}(k)\right| \tag{3}$$

where $\hat{D}(k)$ is the average magnitude of the noise spectrum. The noise estimation is obtained on the assumption that the background noise is locally stationary, so that the noise characteristics computed during the speech pauses are a good approximation to the noise characteristics [2].

The enhanced spectrum obtained using the above relation may contain some negative values due to the errors in estimating the noise spectrum. The simplest solution is half-wave rectification of these values, to ensure a non negative magnitude spectrum. This nonlinear processing of negative values creates small, isolated peaks in the spectrum occurring at random frequency locations in each of the frames. Converted in the time domain, these peaks sound similar to the tones with frequencies that change randomly from frame to frame; that is, tones that are turned on and off at the analysis frame rate. This type of noise is commonly referred as musical noise [3]. The musical noise can be more annoying to the listeners than the original distortion caused by the background noise. Several modifications for the standard spectral subtraction method have been proposed to alleviate the speech distortion introduced by the spectral subtraction process [2].

Boll [1] proposed few modifications such as magnitude averaging, residual noise reduction and additional signal attenuation during non-speech activity, to reduce the effect of musical noise. Berouti *et al.* [3] suggested a method to reduce the musical noise by subtracting an overestimate of the noise spectrum, while preventing the resultant spectral components from going below a preset minimum value. The proposed technique has the following form [3]:

$$\left|\hat{S}(k)\right| = \begin{cases} \left|Y(k)\right| - \alpha\left|\hat{D}(k)\right|, & \left|Y(k)\right| - \alpha\left|\hat{D}(k)\right| > \beta\left|\hat{D}(k)\right| \\ \beta\left|\hat{D}(k)\right|, & \text{otherwise} \end{cases} \tag{4}$$

where α is the over subtraction factor, which is a function of the noisy Signal to Noise Ratio (SNR) and calculated as [3]:

$$\alpha = \alpha_0 - \frac{3}{20}SNR, \quad -5dB \le SNR \le 20dB \tag{5}$$

where α_0 is the desired value of α at 0 dB SNR. Here, SNR is computed as the ratio of the noisy speech power to the estimated noise power. In general, the higher the amount of over subtraction, more will be the attenuation of the stronger components with a low SNR. This prevents musical noise. However, too strong over subtraction will suppress too many components. Therefore, the value of α has to be carefully chosen, in order to prevent both the musical noise and signal distortion [3]. The introduction of spectral floor β prevents the subtraction of spectral components of the enhanced speech spectrum falling below the predefined lower value.

A frequency adaptive subtraction factor based approach is proposed in [4, 5]. The motivation is that, in general, noise may not affect the speech signal uniformly over the whole spectrum. Some frequencies are affected more severely than the others. Accordingly, Lockwood and Boudy [4] proposed the nonlinear spectral subtraction (NSS) method, based on the linear spectral subtraction proposed by Berouti *et al.* [3]. In NSS, the over subtraction factor is frequency dependent in each frame of speech. Larger values are subtracted at frequencies with low SNR levels and smaller values are subtracted at frequencies with high SNR levels. Kamath and Loizou [5] extended this concept and developed a multi-band spectral subtraction method that divides the speech spectrum into N nonoverlapping bands, and the over subtraction factor for each band is calculated independently. The individual frequency bands of the estimated noise spectrum are subtracted from the corresponding bands of the noisy speech spectrum. The performances of the above methods are not satisfactory in adverse environments, particularly when the SNR is very low. The reason is that in very low SNR conditions, it is still difficult to suppress noise without degrading

intelligibility, and without introducing residual noise and speech distortion.

Due to this fact, several perceptual-based approaches are advocated, wherein instead of completely eliminating the musical noise and introducing distortion, the noise is masked taking advantage of the simultaneous masking properties of the auditory system [6, 7]. The masking effect means that a stronger signal can make a weaker signal occurring simultaneously inaudible. If the noise signal is weaker than the speech signal in the same frequency band, then the noise signal is masked by the speech signal [6]. Therefore, we can use less noise subtraction to avoid unnecessary distortion. Accordingly, instead of attempting to remove all noise from the signal, these algorithms attempt to attenuate the noise below the audible threshold [6, 7]. The methods that adopt the masking property of the human auditory system can reduce the effect of residual noise, but the drawback is the large computational effort associated with the subband decomposition and the additional discrete Fourier transform (DFT) analyzer required for psychoacoustic modeling. Even though several improvements have been proposed, spectral subtraction approach is still a subject of many researches on how to increase its performance in terms of minimizing the effect of musical noise and also on making it suitable for non-stationary environments.

2.1.2 MMSE Estimator

In spectral subtraction based methods, there were no specific assumptions made about the distribution of the spectral components of either speech or noise. Ephraim and Malah [8] have proposed a system that utilizes the MMSE criteria using models for the distribution of the spectral components of speech and noise signals. The MMSE-short time spectral amplitude (STSA) estimator for speech enhancement aims to minimize the mean square error between the short time spectral magnitude of the clean and enhanced speech signal. This method assumes that each of the Fourier expansion coefficients of the speech and of the noise process can be modeled as independent, zero-mean, Gaussian random variables [8]. In [9], to incorporate perceptually significant information into [8], the authors proposed a method to minimize the mean square error between the logarithm of the STSA of the clean and enhanced speech. This criterion of optimality gives good results in practice, with a noticeable reduction in musical noise.

The MMSE log-spectral amplitude (MMSE-LSA) estimator for speech enhancement was also proposed by Ephraim and Malah in 1985 [9]. The aim of the authors, in their previous work on MMSE estimation of the STSA, was to enhance the speech by minimizing the error between the STSA of clean speech and enhanced speech. This optimality criterion does not consider any of the nonlinear characteristics observable in human perception [10]. To incorporate perceptually significant information into the algorithm, the authors proposed a method to minimize the mean square error between the logarithm of the STSA of the clean and enhanced speech. That is, the LSA estimator minimizes [2]

$$E\left\{\left(\log_e A_k - \log_e \hat{A}_k\right)^2\right\} \tag{6}$$

where A_k denotes the spectral speech amplitude, and $\hat{A}_k$ is its optimal estimator.

A fundamental assumption made in the MMSE algorithms is that the real and imaginary parts of the clean DFT coefficients can be modeled by a Gaussian distribution. This Gaussian assumption might hold for the DFT coefficients of the noise, typically estimated using relatively short (20-30 ms) duration windows [2]. Based on this observation, a similar optimal MMSE-STSA estimator using non-Gaussian distributions is proposed. In particular, the Gamma [11] or the Laplacian [12] probability distributions are used to model the distributions of the real and imaginary parts of the DFT coefficients.

All MMSE-based methods need the estimate of the *a priori* SNR, the SNR of the k^{th} spectral component of the clean speech signal. Since the knowledge of clean signal is seldom available in practical systems, a decision-directed estimation and maximum likelihood (ML) estimation are used to compute *a priori* SNR [8]. Cappe [13] provided a more detailed analysis on the decision-directed estimation approach and proposed a lower limit to the estimate of the *a priori* SNR, in order to reduce the annoying musical tones. Cohen introduced causal and non-causal recursive estimators for the *a priori* SNR, which take into account the time-frequency correlation of speech signals [14]. The causal *a priori* SNR estimator is closely related to the decision-directed estimator. The non-causal *a priori* SNR estimator employs future spectral measurements to predict better the spectral variances of clean speech. Experimental results show that the non-causal estimator yields a higher improvement in the segmental SNR and lower log-spectral distortion than the decision-directed method and the causal estimator [14]. Even though several improvements have been made in MMSE estimator, the algorithms proposed by the Ephraim and Malah [8,9] are still considered the state of art algorithms for noisy speech enhancement.

2.1.3 Wavelet Denoising

Most of the speech enhancement algorithms are applied in the frequency domain, using short-time Fourier transform (STFT), which allows analyzing nonstationary speech signals. STFT provides a compromise between time resolution and frequency resolution. However, once

the frame length is chosen, the time resolution is the same for all frequency components.

Some of the speech enhancement algorithms are developed using wavelet transform, which provides more flexible time-frequency representation of speech [15]. One popular technique for wavelet-based signal enhancement is the wavelet shrinkage algorithm [16]. Wavelet shrinkage is a simple denoising method based on the thresholding of the wavelet coefficients. The estimated threshold defines the limit between the wavelet coefficients of the noise and those of the target signal. However, it is not always possible to separate the components corresponding to the target signal from those of noise by simple thresholding. For noisy speech, energies of unvoiced segments are comparable to those of noise. Applying thresholding uniformly to all wavelet coefficients not only suppresses additional noise but also some speech components, like unvoiced ones [15]. Consequently, the perceptual quality of the filtered speech is affected. Therefore, the wavelet transform combined with other signal processing tools like Wiener filtering in the wavelet domain and wavelet filter bank have been proposed for speech enhancement [17]. More recently, a number of attempts have been made to use perceptually motivated wavelet decompositions, coupled with various thresholding and estimation techniques [18].

2.2 Temporal Processing Methods

2.2.1 LP Residual Enhancement

Most of the studies on the speech enhancement discussed above focus on enhancement, based on suppression of noise. These methods disturb the spectral balance in speech, resulting in unpleasant distortions in the enhanced speech. Yegnanarayana *et al.* proposed a noisy speech enhancement method by exploiting the characteristics of excitation source signal such as the LP residual [19]. The basic approach for speech enhancement is to identify the high SNR portions in the noisy speech signal, and enhance those portions relative to the low SNR portions, without causing significant distortion in the enhanced speech. The residual signal samples are multiplied with the weight function, and the modified residual is used to excite the time-varying all-pole filter derived from the given noisy speech, to generate the enhanced speech. In this method, enhancement is carried out by the following three steps [19]: (i) identification and enhancement of high SNR regions at the gross level; (ii) identification and enhancement of high SNR regions at the fine level; and, (iii) enhancement of spectral peaks over valleys.

In [20], a speech enhancement algorithm, similar to [19], has been proposed. It differs with the former residual weighting scheme in that the weights on the LP residuals have been derived, based on a constrained optimization criterion. Enhanced speech is obtained by exciting the time-varying all-pole synthesis filter with the enhanced residual. In [21], the authors exploited the use of coherently added Hilbert envelope (HE) for LP residual reconstruction. Large amplitude at the instant of strong excitation, a feature of HE, makes it a good indicator of glottal closure (GC), where an excitation pulse takes place. Therefore, applying HE to the LP residual as a weighting function has the effect of emphasizing the pulse train structure for voiced speech, which leads to an enhanced LP residual signal.

3. Enhancement of Reverberant Speech

Reverberation affects the quality of speech, in which delayed copies of the speech waveform, called echoes are added to the direct speech. Mathematically, this can be expressed as convolution of the speech signal, with room impulse response [22]. That is,

$$z(n) = s(n) * h(n), \tag{7}$$

where s(n) represents the clean speech, h(n) denotes the room impulse response and * symbolizes the convolution operation. The reverberation is completely characterized by the speaker to receiver room impulse response. This can be divided into three segments: Direct sound, early reflections and late reflections. The first sound that is received without reflection is the direct sound. A little later, the sounds which were reflected off one or more surfaces will be received. These reflected sounds are separated from the direct sound, in both time and direction [23]. The reflected sounds form a sound component usually called early reverberation. Early reverberation is actually perceived to reinforce the direct sound and is, therefore, considered useful to speech intelligibility. Late reverberation results from the reflections that arrive with larger delays, after the arrival of the direct sound. They are perceived either as separate echoes or as reverberation and impairing speech intelligibility [23].

Various methods for improving the performance in reverberant environments have been proposed. These methods may also be broadly grouped into *temporal processing* and *spectral processing methods*. The temporal processing methods obtain the enhancement by processing the reverberant speech in time or cepstral domain and spectral domain processing is accomplished in the frequency domain. Besides these categories, there are several *multi-stage algorithms* that have been proposed, which process degraded signal in both time and frequency domains.

3.1 Spectral Processing Methods

The spectral enhancement methods achieve dereverberation by modifying the short-time magnitude spectrum of the reverberant speech. Initially, Flanagan *et al.* [24] proposed a spectral based two microphone approach for

processing reverberant speech. The speech signal from each microphone was separated into several subbands. In each frequency band, the spectral amplitudes of the two signals were compared and the maximum amplitude was selected as the contribution for the reconstructed speech. This method exploits the periodic nature of the spectral distortion of speech caused by simple echo. The two microphones spaced at different locations have echoes of different delays and nulls appear at regular but different intervals in the spectra of the microphone outputs. An algorithm proposed by Allen *et al.* [25] first filters the two individual microphone signals into the frequency band and then the filtered outputs are compensated for the delay differences and added. For each band, the correlation between the two microphone signals is computed. These correlations are used as a gain factor for that band, to suppress the spectral bands with low correlations, in order to remove the reverberation effects. This is done based on the assumption that bands with high levels of coherence contain a strong direct component, whereas bands with low levels of coherence mainly contain reverberation.

Another form of spectral processing method is proposed in [26], using the harmonic structure of speech, based on pitch estimation. In the proposed method, the harmonic part of the speech was extracted by adaptive filtering. Averaging the ratio of DFT of the harmonic part of speech and that of the reverberant part, a dereverberation filter was calculated, which reduced reverberation in both voiced and unvoiced speech segments. The technique is suitable when sufficient number of training utterances are available and the room impulse response does not change significantly.

Recently, a spectral subtraction based spectral processing technique has been developed, to suppress late reverberation effect [23,27]. In spectral subtraction based methods, the impulse response of the room h(n) can be split as shown below:

$$h(n) = \begin{cases} h_{\alpha}(n) & \text{for } 0 \le n \le N_1 \\ h_1(n) & \text{for } N_1 + 1 \le n \le N_i \\ 0 & otherwise \end{cases} \quad (8)$$

where N_i is the length of the impulse response and N_1 is the threshold, which is chosen such that $h_a(n)$ consists of the direct signal and a few early reflections, and $h_l(n)$ consists of all late reflections. The threshold N_1 can be chosen, depending on the application or subjective preference [23].

The spectral subtraction method is developed based on the fact that the early and late impulse components are approximately uncorrelated in the time domain. Therefore, the late reverberant signal can be treated as an additive noise, and thus can be eliminated through spectral subtraction [27]. Accordingly, the short-time power spectral density (PSD) of the reverberant speech signal $S_{zz}(l, k)$ is expressed as [27]:

$$S_{zz}(l, k) = S_{za}(l, k) + S_{zl}(l, k) \quad (9)$$

where $S_{za}(l, k)$ and $S_{zl}(l, k)$ are the PSD of the early and late reverberant components, respectively. Indices l and k refer to the time frame and frequency bin, respectively.

Lebart *et al.* [27] introduced a single channel speech dereverberation method, based on spectral subtraction, to reduce this effect. The method estimates the power spectrum of the reverberation, based on a statistical model of late reverberation, and then subtracts it from the power spectrum of the reverberant speech. In [27], the authors assumed that reverberation time is frequency independent and that the energy related to the direct sound could be ignored. Hence, the authors assume that the signal to reverberation ratio (SRR) of the observed signal is smaller than 0 dB, which limits the use of the proposed solution to a situation in which the source-microphone distance is larger than the critical distance. This issue has been addressed in [23], where the room impulse response model was generalized by considering the direct component and the reflections separately. A novel estimator was derived, which had an advantage over the late reverberant power spectral density estimator proposed by Lebart, if the source-microphone distance was smaller than the critical distance [23].

One of the main problems in spectral subtraction is the nonlinear processing distortion, for example, the musical noise caused by over-subtraction of the reverberation. This distortion degrades the quality of the processed speech. However, there are some well known methods like spectral floor factor, *a priori* SRR estimation techniques available to reduce this nonlinear distortion to a certain level.

3.2 Temporal Processing Methods

3.2.1 Inverse Filtering

Reducing reverberation through inverse filtering is one of the most common approaches.

The basic idea is to pass the reverberant signal through a second filter, which inverts the reverberation process and recovers the original signal. This can be written as

$$s(n)^{*}h(n)^{*}g(n) = \hat{s}(n) \quad (10)$$

where g(n) is the inverse filter impulse response and $\hat{s}(n)$ is the delayed replica of s(n).

There are several well known inverse filtering methods to dereverberate the original signal [22,28,29]. The

challenge in the inverse filtering method is to find the inverse impulse response g(n). The perfect reconstruction of the original signal exists only if the room impulse response function is a minimum phase filter, whose poles and zeros are all inside the unit circle. But in practical case, most room transfer functions are non minimum phase, due to the late energy in the room impulse response and, therefore, inverse filtering based techniques have limited scope in practice [22].

3.2.2 Cepstral Filtering

Oppenhiem *et al.* [30] proposed a single microphone dereverberation approach, using the cepstral filtering technique in which speech is considered as slowly varying in the cepstral domain, with its cepstral components concentrated around the cepstral origin. Whereas, the acoustic impulse response is characterized by the pulses with rapid ripples concentrated far away from the cepstral origin [30]. The complex cepstrum of reverberant speech z(n) can be represented as

$$C_z(n) = C_s(n) + C_h(n) \quad (11)$$

where $C_h(n)$ is the complex cepstrum of the reverberant impulse response and $C_s(n)$ is the complex cepstrum of the speech signal. Therefore, the dereverberation can be achieved by removing the cepstral components corresponding to the impulse response by applying low time lifter in cepstral domain. Also discussed in [30] was an alternative approach, where a cepstral filtering procedure using a comb filter is considered for reducing the reverberation effect.

The cepstral filtering has been successfully applied to the enhancement of speech degraded by simple echoes [30]. Typically, frame based processing is used to calculate the cepstrum of a signal. Since reverberation effects are generally much longer than typical frame lengths, the current frame does not contain all the reverberation effects of the frame, while it also contains reverberation effects from previous frames. Besides, the cepstrum of the clean speech signal C_s (n) and the cepstrum of the acoustic impulse response C_h (n) typically have a large overlap, resulting in signal distortion when using low time filtering. By using an exponential windowing procedure and cepstral averaging in order to identify the room impulse response h(n) before inverse filtering, a significant improvement is possible [31]. However, in practice, single-microphone cepstrum based techniques for dereverberation have a limited performance.

3.2.3 Temporal Envelope Filtering

Various single microphone algorithms are proposed using modulation transfer function (MTF) of speech [32]. According to this method the envelope of the reverberant speech can be approximated by the convolution of the clean speech signal, with the envelope of an acoustic impulse response. Therefore, the problem of enhancement reduces to the deconvolution of the room response envelope and the reconstruction of the speech signal. These methods do not require the impulse response of an environment to be measured [34]. For example, Langhans and Strube [33] proposed an enhancement method, where they appropriately filtered the envelope signals in critical frequency bands based on STFT and linear prediction [34]. They used theoretically derived inverse MTF as high pass filtering to reduce the effect of reverberation. Similarly, Aveandano and Hermansky [32] attempted to recover the energy envelope of the original speech by applying theoretically derived inverse MTF and an optimum filter trained from clean and reverberant speech [34]. To realize this approach, it is assumed that the carrier signal of the speech and the impulse response are white noise. However, these assumptions are not accurate with regard to real speech and reverberation [34]. Therefore, this approach has not yet achieved high quality dereverberation.

3.2.4 LP Residual Enhancement

Yegnanarayana and Murthy developed a reverberant speech enhancement system by manipulating excitation source information based on the residual characteristics of speech [35]. Manipulation of the residual signal is more appropriate than the manipulation of speech signal, especially for short segments, as the residual signal samples are generally less correlated than the speech samples. On the other hand, for manipulation of the speech signal directly, the choice of the size and shape of the window may affect the results significantly. The processing method involves identifying and manipulating the linear prediction residual signal in different regions of the reverberant speech signal, namely, regions in which there is high SRR, low SRR and reverberant component only. 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 high SRR and low SRR regions of the reverberant speech and the fine level weight function is derived to enhance the instants of significant excitation of original signal. In [35], gross level identification is done using the entropy of the distribution of the samples in the LP residual signal and fine level identification is done using the normalized prediction error. The authors also observed that there was a reduction in the flatness of the spectral

envelope, owing to reverberation. Thus, the LP coefficients are manipulated to increase the spectral flatness. Finally, the enhanced speech signal is resynthesized from the processed LP residual signal and the coefficients. In [36], the authors proposed a multi-channel speech enhancement technique by exploiting the features of the excitation source in speech production. The Hilbert envelope (HE) of LP residual was used to derive the information of the strength of excitation. A weight function was derived by coherently combining the delay compensated HEs of the LP residual signals from the different microphones. The enhanced speech was again obtained by exciting the time-varying all-pole filter with the LP residual modified by the weight function.

In [37], the authors presented a spatiotemporal averaging method for the enhancement of reverberant speech. The basis was that the waveform of the LP residual between adjacent larynx-cycles varied slowly, so that each such cycle could be replaced by an average of itself and its nearest neighboring cycle. The averaging resulted in the suppression of spurious peaks in the LP residual, caused by room reverberation. Finally, a speech signal with reduced reverberation was synthesized with the enhanced LP residual.

Most of the LP residual techniques rely on the important assumption that the calculated LP coefficients of the all-pole filter are unaffected by the multi-path reflections of the room. Gaubitch and Naylor showed that this assumption holds only in a spatially averaged sense, and that it cannot be guaranteed at a single point in space for a given room [38]. Recently, Gaubitch *et al.* used statistical room acoustic theory for the analysis of the auto regressive (AR) modeling of reverberant speech [39]. They investigated and showed that proper calculation of the LP coefficients, i.e., using spatially averaged LP coefficients, improved the quality of LP residual enhancement techniques.

3.3 Multi-Stage Algorithms

During the last decade, several multi-stage algorithms were proposed for the enhancement of reverberant speech. In [26], Nakatani and Miyoshi proposed a system capable of blind dereverberation of one microphone speech, by employing the harmonic structure of speech. In this system, a sinusoidal representation was used to approximate the direct sound in the reverberant environments and adaptive harmonic filters were first employed to estimate the voiced clean speech from the reverberant speech signal. This estimation was then used to derive a dereverberation filter. This method, however, requires accurate estimation of the fundamental frequency from the reverberant speech. Wu and Wang [40] proposed a two-stage model to enhance reverberant speech. In the first stage, an inverse filter of the room impulse response was estimated, to increase the SRR by maximizing the kurtosis of the LP residual. In the second stage, long term reverberation effects were removed using the spectral subtraction approach. In [41], a hybrid dereverberation method was proposed, which combined correlation based blind deconvolution and modified spectral subtraction to suppress the tail of reverberation and improve the processed speech quality. Inverse filtering reduced the early reflection that constitutes most of the power of the reverberation. Then, the modified spectral subtraction suppressed the tail of the inverse-filtered speech.

4. Enhancement of Multi-speaker Speech

A source of degradation, which is more difficult to handle owing to the speech of a competing speaker, is popularly known as cocktail party effect. This case is difficult for enhancement because the degrading signal too has the characteristics of speech, which makes it difficult to distinguish it from the desired signal. This is primarily because: (i) The pitch and formants of different talkers may cross or overlap (ii) The number of talkers is usually not known, and (iii) Each talker's amplitudes vary within the utterance.

Several approaches have been proposed in the literature to process speech degraded by speech of competing speaker. Most of these methods may be broadly grouped into three categories – blind source separation (BSS) using independent component analysis (ICA), computational auditory scene analysis (CASA), and speech-specific approaches (SSA). Here, the first two categories are well known to the speech processing community. Speech processing in a multi-speaker environment is also attempted by the speech processing community, with an aim to use available speech-specific knowledge, like short time spectrum analysis, gross characteristics of excitation (voiced and unvoiced features), cepstrum, fundamental frequency, segmentation and masking, in time-frequency planes for separation. We group them as speech-specific approaches. Depending on the number of microphones used for collecting speech, these methods may be further classified into single and multi-channel cases. In single channel case, speech is collected over a single microphone, and the objective is to process multi-speaker speech to emphasize desired speaker's speech. This approach is more commonly termed as co-channel speaker separation [42]. In multi-channel case, speech is collected simultaneously over several (two or more) spatially distributed microphones. Signals from all the microphones are processed to enhance the speech of one or more speakers. Separation of speech signals can be done effectively, if the speech signals are collected simultaneously over two or more microphones. This is mainly because multi-channel methods exploit the spatial diversity resulting from the fact that

desired and undesired speakers are in practice, located at different points in space. Similar to noisy speech and reverberant speech, these methods can also be grouped into temporal and spectral processing methods.

4.1 Spectral Processing Methods

4.1.1 Speech-Specific Approaches

Initial work in co-channel speaker separation evolved from speech enhancement algorithms designed for separating voiced speech from background noise, given a pitch estimate from the target talker. Pearson [42] proposed a harmonic selection method for co-channel speech separation. In this method, first the spectral peaks are identified from the windowed mixed speech spectrum. The peaks are accumulated in a table that is used to construct a histogram. The pitch (F0) of a first speaker is determined from the histogram and the F0 of the second speaker is obtained by removing the harmonics belonging to the first speaker from the peak table and repeating the histogram calculation for remaining peaks. The speech of each speaker is then resynthesized by taking inverse discrete Fourier transform (IDFT) of separated pitch and harmonics. Morgan *et al.* [43] proposed a harmonic enhancement and suppression algorithm for separating the two speakers. The idea was to recover the stronger speaker's speech by enhancing his/her harmonics and formants, given a multi resolution pitch estimate. The weaker speaker's speech is then obtained from the residual signal created, when the harmonics and formants of the stronger talker are suppressed. When there are more than two talkers in the co-channel signal, only the stronger speaker can be separated, and the separation is predicated on the basis that the speaker is always stronger and voiced.

Sinusoidal modeling of speech is also suggested to obtain the co-channel speaker separation [44]. The enhancement is achieved by synthesizing a waveform from the sine waves of desired speaker with the help of *a priori* sine wave frequencies or *a priori* pitch contour and least square estimation technique. The basic requirement of all these methods is that the voices to be separated must be periodic. Generally the separation of unvoiced speech is more difficult, as compared to voiced speech. This is mainly because of two reasons. Firstly, unvoiced speech lacks harmonic structure and is often acoustically noise-like. Secondly, the energy of unvoiced speech is usually much weaker than that of voiced speech. However, by the nature of speech production, most of the speech produced is of the voiced type, and, hence, nearly all the information is perceived from the voiced sounds itself.

4.1.2 CASA Methods

While speech enhancement using signal processing methods with satisfactory performance remains a challenge, the natural ability to enhance sounds of interest selectively by the human auditory system inspired researchers to approach this issue in a different way. In 1990, Bregman proposed the concept of auditory scene analysis (ASA) to segregate acoustic signal into streams, which correspond to different sources [45]. A typical ASA system generally consists of two main stages: Segmentation (analysis) and grouping (synthesis). In the first stage, the mixture sound is segmented into the time-frequency cells. Segmentation is performed using either the STFT or the gammatone filter bank [46]. The segments are then grouped, based on the cues that are mainly onset and offset, on harmonicity, and on position cues [47]. This ASA account has inspired a series of computational ASA (CASA) systems for sound segregation [46]. A main advantage of CASA is that it does not make strong assumptions about interference. Generally, a typical CASA system contains four stages: peripheral analysis, feature extraction, segmentation, and grouping [47]. The peripheral processing decomposes the auditory scene into a time-frequency (T-F) representation via bandpass filtering and time windowing. The second stage extracts auditory features corresponding to ASA cues. In segmentation and grouping, the system generates segments for both target and interference, and groups the segments originating from the target into a target stream. Finally, the waveform of the segregated target is synthesized from the target stream [47]. The techniques based on CASA suffer from two problems. First, these techniques are not able to separate unvoiced segments and almost in all reported results, one or both underlying signals are fully voiced [48]. Second, the vocal-tract related filter characteristics are not included in the discriminative cues for separation. In other words, in CASA techniques, the role of the excitation signal is more important than the vocal tract shape [49].

4.2 Temporal Processing Methods

4.2.1 LP Residual Enhancement

A method for processing speech from a multi-speaker environment, using excitation source information, is proposed by the authors in [50]. The speech of each speaker is enhanced with respect to the speech of the other, by performing the relative emphasis of speech signal around each instant of significant excitation of the desired speaker. The relative emphasis is achieved by giving a larger weight to the LP residual samples in the region, around the instants of significant excitation, and lower weight to the samples in the other regions [50].

The temporal processing approach proposed in [50] composes of following steps: (i) Identification of instants of significant excitation for determining the short high energy regions corresponding to each speaker, (ii) Classification of extracted instants into two speaker classes, (iii) Weighting

the LP residual to enhance the excitation characteristics of desired speaker, and (iv) Synthesizing the enhanced speech by exciting the time-varying all-pole filter with the LP residual modified by the weight function. The HE of LP residual is used as a representation for the sequence of impulses corresponding to the instants of significant excitation of the vocal tract system [50]. When these sequences are added coherently, using the knowledge of the time-delay of each speaker, the strengths of the excitation of the desired speaker are enhanced, relative to the strengths of excitation of other speakers. From the coherently added sequence of impulses, a weight function is derived, which is used to derive a modified excitation signal. This modified excitation signal is used to synthesize speech using the vocal-tract system characteristics derived from the degraded speech.

4.2.2 Cepstral Processing

Stubbs and Summerfield [51] compared the harmonic selection procedure suggested by Parsons [42], with the cepstral transformation of speech. The cepstral transformation maps the spectral envelope to a region near the origin of the cepstral domain, and maps the harmonic excitation to a position well separated from the origin. For voiced speech, the harmonic excitation was simply an impulse with cepstrum quefrency equal to the pitch period. If the pitch peak in the cepstrum was attenuated, the harmonic excitation was reduced. The success of this filtering operation usually requires one voice to be stronger than the interfering voice. Therefore, speech separation not only depends on the processing method used but also on the nature of the degraded signal. The more separated the pitch and harmonics of each talker, the better the results to be expected.

4.3 BSS and ICA Methods

Recently, blind source separation (BSS) by independent component analysis (ICA) has received great attention. Blind separation of instantaneous mixture is achieved by the ICA, which aims at decomposing the multivariate data into a linear sum of independent components [52]. The goal in BSS is to recover a set of independent sources, given only a set of sensor observations that are generated from the individual source signals, through an unknown linear mixing process [53]. In BSS research, there are two important problems that are generally considered: Instantaneous BSS and convolutive BSS. In the case of instantaneous BSS, signals are mixed instantaneously. However, in a practical environment, signals are always mixed in a convolutive manner, because of the reverberation effects.

The BSS technique in speech signal separation was first attempted by Cardoso [54] and Jutten [55], using the principle of statistical independence of the sources [53]. Blind separation of multiple speakers was attempted where the coefficients of the finite impulse response (FIR) filters were used to represent the linear mixing of the sources. These algorithms were based on the higher order statistics of the signal's mutual independence measure among the independent components. Later, numerous approaches have been presented using ICA in BSS for speech separation [56-58]. Even though a variety of algorithms have been proposed, all ICA algorithms are fundamentally similar. The main difference between the different ICA algorithms is the numerical algorithm used for measuring the signal independence. The basic ICA approach uses the following linear model [59]:

$$X = AS \tag{12}$$

where the vector S represents m independent sources, the matrix A represents the linear mixing of the sources, and the vector X is composed of m observed signals. The idea of the ICA is to recover the original sources by assuming that they are statistically independent. The independence assumption means that the joint PDF is the product of the densities for all sources.

$$P(s) = \prod_i p(s_i) \tag{13}$$

where $p(s_i)$ is the PDF of source i and P(S) is the joint probability density function.

BSS using ICA achieves near perfect reconstruction of independent sources, in case of synthetic mixture of speech signals. However, when applied to a mixture of speech signals collected from real acoustic environments, the performance degrades severally due to the effect of reverberation and background noise.

Several methods have been proposed and are being proposed in the framework of BSS using ICA, to improve the performance in real acoustic environments [60]. In spite of these sustained efforts, the performance is still not satisfactory and there is a belief that using more *a priori* information about speech may help to improve the performance.

ICA based algorithms for separation of speech signal have been developed in the domains of both time and frequency. The time domain approach achieves good separation results, once the algorithm converges. However, these methods suffer from a large computational load, to compute convolution of long filters. The frequency domain BSS has a great advantage that the convolution in the time domain becomes multiplication in the frequency domain and it can be easily implemented using FFT with less number of computations. However, the problems with frequency domain approach are indeterminacy of scaling and permutation.

5. Summary and Conclusion

In this paper a brief review of various temporal and spectral processing approaches for the enhancement of degraded speech has been made. The review has mainly focused from the temporal and spectral processing perspective. Various temporal and spectral processing approaches for the enhancement of noisy speech, reverberant speech and multi-speaker speech have been discussed.

As it can be observed from the above discussion, the underlying principle of processing degraded speech is different in each domain of processing. Considering STFT based spectral processing, the focus of most of the spectral processing methods for speech enhancement is on the estimation (i.e., spectral characteristics of background noise, late reverberation, interfering speaker) and suppression of the degradation rather than enhancement of the characteristics of the speech signal. Information about the degradation needs to be continuously estimated, particularly in non-stationary environments wherein degradation characteristics are constantly changing. Alternatively, the temporal processing methods that use the characteristics of excitation source information primarily aim at emphasizing the high SNR/SRR regions of degraded speech signal. Therefore, no explicit knowledge of characteristics of degradation is required.

The limitation of the temporal processing methods is that the level of removal of degradation achieved may not be significant, as in the case of spectral based methods. These two approaches may, therefore, be effectively combined by exploiting their merits and aiming to minimize the demerits. For instance, in case of noisy speech, the difficulty in estimating degradation in the highly non-stationary environment for spectral processing may be carried out by using the gross weight function derived from the temporal processing as a voice activity detector to identify nonspeech regions. Similarly, in the case of reverberant speech, the temporal processing methods enhance the speech-specific features of high SRR regions in the temporal domain. The spectral subtraction based spectral processing methods reduce the late reverberation by estimating and subtracting the late reverberant spectrum from the degraded speech spectrum. The combination of these methods effectively enhances reverberant speech at high SRR regions and eliminates late reverberation. Further, each domain of processing uses independent speech-specific features for processing; that is, excitation source features for temporal processing and vocal-tract based spectral features for spectral processing. Thus, it may be possible to combine these two processing approaches for exploiting their independent nature of processing degraded speech.

Therefore, in general, the combination of temporal and spectral processing methods may lead to speech enhancement methods that are more effective and robust, when compared to any one of them. In future, combined TSP methods can be developed, which may result in improved performance, as compared to the individual temporal or spectral processing methods.

References

1. S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal process.*, vol. ASSP-27, pp. 113-20, Apr. 1979.
2. P. C. Loizou, *Speech Enhancement: Theory and Practice*, 1st ed. Boca Raton, FL.: CRC, 2007.
3. M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal process.*, Apr. 1979, pp. 208-11.
4. P. Lockwood, and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars," *Speech Communication*, vol. 11, no. 2-3, pp. 215-28, 1992.
5. S. Kamath, and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal process.*, Orlando, USA, May 2002.
6. Ming-Chan You, Cheng-Yi Mao, and Jeen-Shing Wang, "Recursive Parametric Spectral Subtraction Algorithm for Speech Enhancement," *Communications in Computer and Information Sciences*, vol. 2, pp. 826-35, 2007.
7. N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio process.*, vol. 7, pp. 126-37, Mar. 1999.
8. Y. Ephraim, and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal process.*, vol. ASSP-32, pp. 1109-21, Dec. 1984.
9. Y. Ephraim, and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal process.*, vol. ASSP-33, pp. 443-5, Apr. 1985.
10. B. J. Shannon, "Speech recognition and enhancement using autocorrelation domain processing," Ph.D. dissertation, School of engineering, Griffith University, Brisbane, Australia, Aug. 2006.
11. M. Marzinzik, and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Trans. Speech Audio process.*, vol. 10, pp. 109-18, Feb. 2002.
12. B. Chen, and P. C. Loizou, "A Laplacian-based MMSE estimator for speech enhancement," *Speech Communication*, vol. 49, pp. 134-43, Feb. 2007.
13. O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio process.*, vol. 2, pp. 345-9, Apr. 1994.
14. I. Cohen, "Speech enhancement using super-Gaussian speech models and noncausal a priori SNR estimation," *Speech Communication*, vol. 47, pp. 336-50, Nov. 2005.
15. H. Tasmaz, and E. Ercelebi, "Speech enhancement based on undecimated wavelet packet-perceptual filterbanks and MMSE-STSA estimation in various noise environments," *Digital Signal process.*, vol. 18, no. 5, pp. 797-812, Sep. 2008.
16. D. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Information Theory*, vol. 41, no. 3, pp. 613-27, May 1995.
17. M. K. Hasan, S. Salahuddin, and M. R. Khan, "Reducing signal-bias from mad estimated noise level for dct speech enhancement," *Signal Process.*, vol. 84, no. 1, pp. 151-62, 2004.
18. J.-H. Chang, S. Gazor, N. S. Kim, and S. K. Mitra, "Multiple statistical models for soft decision in noisy speech enhancement," *Pattern Recognition*, vol. 40, pp. 1123-34, Mar. 2007.

19. B. Yegnanarayana, C. Avendano, H. Hermansky, and P. Satyanarayana Murthy, "Speech enhancement using linear prediction residual," *Speech Communication*, vol. 28, pp. 25-42, May 1999.
20. W. Jin, and M. S. Scordilis, "Speech enhancement by residual domain constrained optimization," *Speech Communication*, vol. 48, pp. 1349-64, Oct. 2006.
21. B. Yegnanarayana, S. R. Mahadeva Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal process.*, vol. 1, Orlando, USA, 2002, pp. I-541-4.
22. S. Neely, and J. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Am.*, vol. 66, pp. 165-9, 1979.
23. E. Habets, "Single-and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Universiteit Eindhoven, The Netherlands, Jun. 2007, http://alexandria.tue.nl/extra2/200710970.pdf.
24. J. Flanagan, and R. Lummis, "Signal processing to reduce multipath distortion in small rooms," *J. Acoust. Soc. Am.*, vol. 47, pp. 1475-81, 1970.
25. J. Allen, D. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Am.*, vol. 62, pp. 912-5, 1977.
26. T. Nakatani, and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal process.*, vol. 1, Hong Kong, China PR, Apr. 2003, pp. 92-5.
27. K. Lebart, and J. Boucher, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoustica*, vol. 87, pp. 359-66, 2001.
28. M. Miyoshi, and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal process.*, vol. ASSP-36, pp. 145-52, Feb. 1988.
29. B. Gillespie, H. Malvar, and D. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptive filltering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal process.*, vol. 6, Salt Lake City, USA, 2001, pp. 3701-4.
30. A. Oppenheim, R. Schafer, and J. T.G. Stockham, "Nonlinear filtering of multiplied and convolved signals," *Proc. IEEE*, vol. 56, pp. 1264-91, Aug. 1968.
31. M. Tohyama, R. Lyon, and T. Koike, "Source waveform recovery in a reverberant space by cepstrum dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal process.*, vol. 1, Minneapolis, USA, Apr. 1993, pp. 157-60.
32. C. Avendano, and H. Hermansky, "Study on the dereverberation of speech based on temporal envelope filtering," in *Proc. Fourth Int. Conf. Spoken Language*, vol. 2, Oct. 1996, pp. 889-92.
33. T. Langhans, and H. W. Strube, "Speech enhancement by nonlinear multiband envelope filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal process.*, vol. 1, May 1982, pp. 156-9.
34. Masashi Unoki, Masakazu Furukawa, Keigo Sakata, and Masato Akagi, "An improved method based on the MTF concept for restoring the power envelope from a reverberant signal," *J. Acoustical Science and Technology*, vol. 25, no. 4, pp. 232-42, 2004.
35. B. Yegnanarayana, and P. Satyanarayana Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech Audio process.*, vol. 8, pp. 267-81, May 2000.
36. B. Yegnanarayana, S. R. M. Prasanna, R. Duraiswami, and D. Zotkin, "Processing of reverberant speech for time-delay estimation," *IEEE Trans. Speech Audio process.*, vol. 13, pp. 1110-8, Nov. 2005.
37. N. Gaubitch, and P. Naylor, "Spatiotemporal averagingmethod for enhancement of reverberant speech," in *Proc. 15th Inter. Conf. Digital Signal process.*, Cardiff, Wales, UK, Jul. 2007, pp. 607-10.
38. N. Gaubitch, P. Naylor, and D. Ward, "On the use of linear prediction for dereverberation of speech," in *Proc. Int. Workshop Acoust., Echo Noise Control*, Sep. 2003.
39. N. D. Gaubitch, D. B. Ward, and P. A. Naylor, "Statistical analysis of the autoregressive modeling of reverberant speech," *J. Acoust. Soc. Am.*, vol. 120, pp. 4031-9, Dec. 2006.
40. M. Wu, and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech, Language process.*, vol. 14, pp. 774-84, May 2006.
41. K. Furuya, and A. Kataoka, "Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction," *IEEE Trans. Audio, Speech, Language process.*, vol. 15, pp. 1579-91, Jul. 2007.
42. T. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Am.*, vol. 60, pp. 911-8, Oct. 1976.
43. D. Morgan, E. George, L. Lee, and S. Kay, "Cochannel speaker separation by harmonic enhancement and suppression," *IEEE Trans. Speech Audio process.*, vol. 5, pp. 407-24, Sep. 1997.
44. T. Quatieri, and R. Danisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," *IEEE Trans. Acoust., Speech, Signal process.*, vol. ASSP-38, pp. 56-69, Jan. 1990.
45. A. S. Bregman, *Auditory scene analysis*. Cambridge, MA: MIT Press, 1990.
46. D. Wang, and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.
47. G. Hu, and D. Wang, "An auditory scene analysis approach to monaural speech segregation," in *Topics in Acoustic Echo and Noise Control*, I. H. E. and S. G, Eds. Springer, Heidelberg, 2006, pp. 485-515.
48. G. Hu, and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Networks*, vol. 15, no. 5, pp. 1135-50, Sep. 2004.
49. M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation," *EURASIP J. Audio Speech Music process.*, vol. 2007, Article ID 84186, 15 pages, 2007. doi:10.1155/2007/84186.
50. B. Yegnanarayana, S. R. M. Prasanna, and M. Mathew, "Enhancement of speech in multispeaker environment," in *Proc. European Conf. Speech process., Technology*, Geneva, Switzerland, 2003, pp. 581-4.
51. R. J. Stubbs, and Q. Summerfield, "Algorithms for separating the speech of interfering talkers: Evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 87, no. 1, pp. 359-72, 1990.
52. P. Comon, "Independent component analysis, a new concept?" *Signal process.*, vol. 36, no. 3, pp. 287-314, 1994.
53. J. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 86, pp. 2009-25, 1998.
54. J.-F. Cardoso, "Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal process.*, vol. 5, Apr. 1990, pp. 2655-8.
55. C. Jutten, and J. Herault, "Blind separation of sources, part 1: An adaptive algorithm based on neuromimetic architecture," *Signal process.*, vol. 24, no. 1, pp. 1-10, 1991.
56. F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "Combined approach of array processing and independent component analysis for blind separation of acoustic signals," *IEEE Trans. Speech Audio process.*, vol. 11, no. 3, pp. 204-15, May 2003.
57. Z. Koldovsky, and P. Tichavsky, "Time-domain blind audio source separation using advanced ICA methods," in *Proc. INTERSPEECH 2007*, Antwerp, Belgium, Aug. 2007, pp. 27-31.
58. N. Das, A. Routray, and P. K. Dash, "ICA methods for blind source separation of instantaneous mixtures: A case study," *Neural Information process. Letters and Reviews*, vol. 11, no. 11, pp. 225-46, Nov. 2007.
59. A. Hyv¨arinen, and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411-30, 2000.
60. S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Speech Audio process.*, vol. 11, no. 2, pp. 109-16, Mar. 2003.

AUTHORS

P. Krishnamoorthy was born in Tamil Nadu, India, in 1980. He received the B.E. degree in electrical and electronics engineering from Thiagarajar College of Engineering, Madurai, India, in 2001 and the M.Tech. degree in applied electronics from P.S.G.College of Technology, Coimbatore, India, in 2003. He is currently pursuing the Ph.D. degree in electronics and communication engineering at Indian Institute of Technology Guwahati, India.

His research interests are in digital signal processing and speech and audio processing.

E-mail: pkm@iitg.ernet.in

S. R. Mahadeva Prasanna was born in India in 1971. He received the B.E. degree in electronics engineering from Sri Siddartha Institute of Technology, Bangalore University, Bangalore, India, in 1994, the M.Tech. degree in industrial electronics from the National Institute of Technology, Surathkal, India, in 1997, and the Ph.D. degree in computer science and engineering from the Indian Institute of Technology Madras, Chennai, India, in 2004. He is currently an Associate Professor in the Department of Electronics and Communication Engineering, Indian Institute of Technology, Guwahati.

His research interests are in speech and signal processing, application of AI tools for pattern recognition tasks in speech, and signal processing.

E-mail: prasanna@iitg.ernet.in

DOI: 10.4103/0256-4602.49103; Paper No TR 50_08;