ELSEVIER

Speech Communication 28 (1999) 25–42

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Speech enhancement using linear prediction residual

B. Yegnanarayana [a,*], Carlos Avendano [b], Hynek Hermansky [b], P. Satyanarayana Murthy [c]

[a] *Department of Computer Science and Engineering, Indian Institute of Technology, Madras 600 036, India*
[b] *Department of Electrical Engineering, Oregon Graduate Institute of Science and Technology, Portland, USA*
[c] *Department of Electrical Engineering, Indian Institute of Technology, Madras 600 036, India*

Received 2 July 1997; received in revised form 12 May 1998

Abstract

In this paper we propose a method for enhancement of speech in the presence of additive noise. The objective is to selectively enhance the high signal-to-noise ratio (SNR) regions in the noisy speech in the temporal and spectral domains, without causing significant distortion in the resulting enhanced speech. This is proposed to be done at three different levels. (a) At the gross level, by identifying the regions of speech and noise in the temporal domain. (b) At the finer level, by identifying the regions of high and low SNR portions in the noisy speech. (c) At the short-time spectrum level, by enhancing the spectral peaks over spectral valleys. The basis for the proposed approach is to analyze linear prediction (LP) residual signal in short (1–2 ms) segments to determine whether a segment belongs to a noise region or speech region. In the speech regions the inverse spectral flatness factor is significantly higher than in the noisy regions. The LP residual signal enables us to deal with short segments of data due to uncorrelatedness of the samples. Processing of noisy speech for enhancement involves mostly weighting the LP residual signal samples. The weighted residual signal samples are used to excite the time-varying all-pole filter to produce enhanced speech. As the additive noise level in the speech signal is increased, the quality of the resulting enhanced speech decreases progressively due to loss of speech information in the low SNR, high noise regions. Thus the degradation in performance of enhancement is graceful as the overall SNR of the noisy speech is decreased. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Speech enhancement; Linear prediction residual signal

1. Introduction

Speech signal collected under normal environmental conditions is usually degraded due to noise and distortion. Performance of speech systems depends critically on the effect of these environmental conditions on the parameters and features extracted from the speech signal (Deller et al., 1993; Rose et al., 1994; Junqua and Haton, 1996). The quality of the recorded speech is also affected significantly due to noise and distortion. Enhancement of speech is normally required to reduce annoyance due to noise. The focus of study in this paper is speech enhancement in additive noise.

Several approaches were studied for speech enhancement in additive noise (Boll, 1979; Gibson et al., 1991; Cheng and O'Shaughnessy, 1991; Ephraim, 1992; Ephraim and Van Trees, 1995; Le Bouquin, 1996; Lee and Shirai, 1996). Many of these studies have focussed on enhancement based

[*] Corresponding author. Tel.: +91-44-2354591; fax: +91-44-2350509; e-mail: yegna@iitm.ernet.in

PII: S 0 1 6 7 - 6 3 9 3 (9 8) 0 0 0 7 0 - 3

on attempts to suppress noise (Boll, 1979; Ephraim and Van Trees, 1995; Le Bouquin, 1996). In order to suppress noise the characteristics of noise are estimated from the regions containing predominantly noise. Therefore for suppressing noise it is necessary to identify the noise regions. Subtraction of noise from noisy speech is usually performed in the spectral domain. Methods based on spectral subtraction disturb the spectral balance in speech, resulting in unpleasant distortion in the enhanced speech. Speech enhancement has also been accomplished by modifying the temporal contours of the parameters or features, like spectral band energies (Hermansky and Morgan, 1994; Avendaño et al., 1996). The technique uses data-dependent filters that reduce the random fluctuations in the parameter contours caused by noise, and thus enhances the characteristics of speech. The parameters of speech are usually related to short-time spectra. Therefore modification of the temporal variations of the spectral features may sometimes introduce unnatural spectral changes which are perceived as distortion in the enhanced speech.

Methods for speech enhancement have also been developed based on extraction of parameters from noisy speech, and synthesizing speech from these parameters (Yegnanarayana and Ramachandran, 1992). All-pole modeling of degraded speech is one such method (Lim and Oppenheim, 1978). In the all-pole modeling, if wrong peaks are extracted, then these peaks may get enhanced. Temporal sequence of these peaks also produces discontinuities in the contours of the spectral peaks when compared with the smooth contours encountered in natural speech.

Methods of speech enhancement seem to depend generally on modification of the short-time spectral envelope. If there are errors in extracting the features of a spectral envelope, or if errors are introduced in the spectral envelope due to modification of the temporal contours of the spectral features, the resulting speech may produce unnatural audible distortion.

Methods for speech enhancement have also been suggested based on the periodicity due to pitch (Erell and Weintraub, 1994). Noise samples in successive glottal cycles are uncorrelated. On the other hand, the characteristics of the vocal tract system are highly correlated due to slow movement of the articulators. These methods for enhancement of speech depend critically on the estimation of pitch from the noisy speech signal. Also, synthetic excitation signal is used for producing speech in these cases. Hence the quality of speech will be poor, even though the effects of noise are reduced.

Several suprasegmental parameters such as pitch contours and syllabic durations are robust features. But these features are not useful for enhancement, since for generating the enhanced speech signal one needs both the spectral envelope and excitation for each (short-time) analysis frame.

In many of the above mentioned methods, no attempt has been made to explore the characteristics of the source signal for enhancement. The primary reason for this is that, in the source signal such as the linear prediction (LP) residual signal the samples are uncorrelated and hence the residual samples are more like noise than like a signal. Thus the residual signal is not expected to have any features useful for speech enhancement. We show in this paper that features of the residual error signal can be exploited for enhancement of speech in the presence of additive noise.

In the next section we discuss the scope of study in this paper. We also discuss the characteristics of noisy speech which form the basis for the proposed approach for speech enhancement. In Section 3, we develop a method for speech enhancement based on the characteristics of the LP residual signal. We propose enhancement at three levels, each level providing improvement of some feature of speech in the noisy signal. In Section 4, we discuss application of the proposed method for different types of additive noise. We also discuss the performance and limitations of the proposed approach.

2. Basis for the proposed method of speech enhancement

Human beings perceive speech by capturing some features from the high signal-to-noise ratio

(SNR) regions in the spectral and temporal domains, and then extrapolating the features in the low SNR regions (Cooper, 1980). Therefore speech enhancement should primarily aim at highlighting the high SNR regions relative to the low SNR regions. Lowering the signal levels in the low SNR regions relative to the signal levels in the high SNR regions may help in reducing the annoyance due to noise without losing the information. The relative emphasis of the features in the high SNR regions over the features in the low SNR regions should be accomplished without causing distortion in speech. Otherwise the enhancement may cause annoyance of a type different from that due to additive noise. The objective in this paper is to study the enhancement produced due to modification of the characteristics of the source and system components of speech production in the signal.

2.1. Effects of noise on the speech signal

Before we proceed to discuss our approach, let us briefly review some characteristics of noisy speech. Speech signal has a large (30–60 dB) dynamic range in the temporal and spectral domains. For example, in the temporal domain some sounds have low signal energy, especially during the release of stop sounds and in the steady nasal sounds. Speech signal energy level is also low prior to the release of a stop sound and also in some fricative sounds. Even within a glottal cycle of a voiced speech signal the energy of the signal is usually higher only in the vicinity of the major excitation of the vocal tract system, which is the instant of glottal closure in each glottal cycle (Ananthapadmanabha and Yegnanarayana, 1979). This is due to damped sinusoidal nature of the impulse response of the vocal tract system. Even in the frequency domain the spectral levels of large amplitude formants are typically much higher (20–30 dB) than the low amplitude formants. The spectral envelope also decreases by 12–18 dB per octave due to glottal roll-off (Fant, 1993). For a given additive noise, the SNR varies as a function of frequency in the spectral domain. Thus the SNR is different in different segments of speech in both time and frequency domains. Fig. 1(c) shows the SNR of a speech utterance as a function of time, where the overall SNR is 10 dB. The noisy speech signal (Fig. 1(b)) is generated by adding white Gaussian noise to the clean speech signal shown in Fig. 1(a). The SNR is computed for each frame of duration 20 ms with an overlap of 10 ms.

Typically, the correlation between noise samples is low, and speech samples are correlated. The envelope of the speech spectrum will be less flat due to formant structure and glottal roll-off compared to the noise spectrum. Additive noise increases the spectral flatness of speech. The spectral envelope becomes more flat in the low SNR portions of the spectrum. As the noise level increases, the weaker spectral features and the low energy signal features will be progressively submerged in the noise. The proposal in this paper 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. Note that, from human perception point of view, some background noise is tolerable, but not the distortion caused by the artifacts of processing.

2.2. Approach for speech enhancement

In this section we present the proposed approach for speech enhancement. We attempt to emphasize the residual signal in the regions around the glottal closure in the voiced speech segments and reduce the energy levels of the residual signal in the silence regions. By exciting the time-varying all-pole filter (derived from the noisy speech) with the modified residual signal, one can produce enhanced speech without causing significant distortion.

Let $\boldsymbol{x} = [x_t, x_{t+1}, \ldots, x_{t+N-1}]^{\mathrm{T}}$ be a frame of N samples of the signal corrupted by additive random noise. The characteristics of the signal are assumed to be stationary within the frame. We can write $\boldsymbol{x}$ as

$$\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{n}, \tag{1}$$

where

$$\boldsymbol{s} = [s_t, s_{t+1}, \ldots, s_{t+N-1}]^{\mathrm{T}}$$

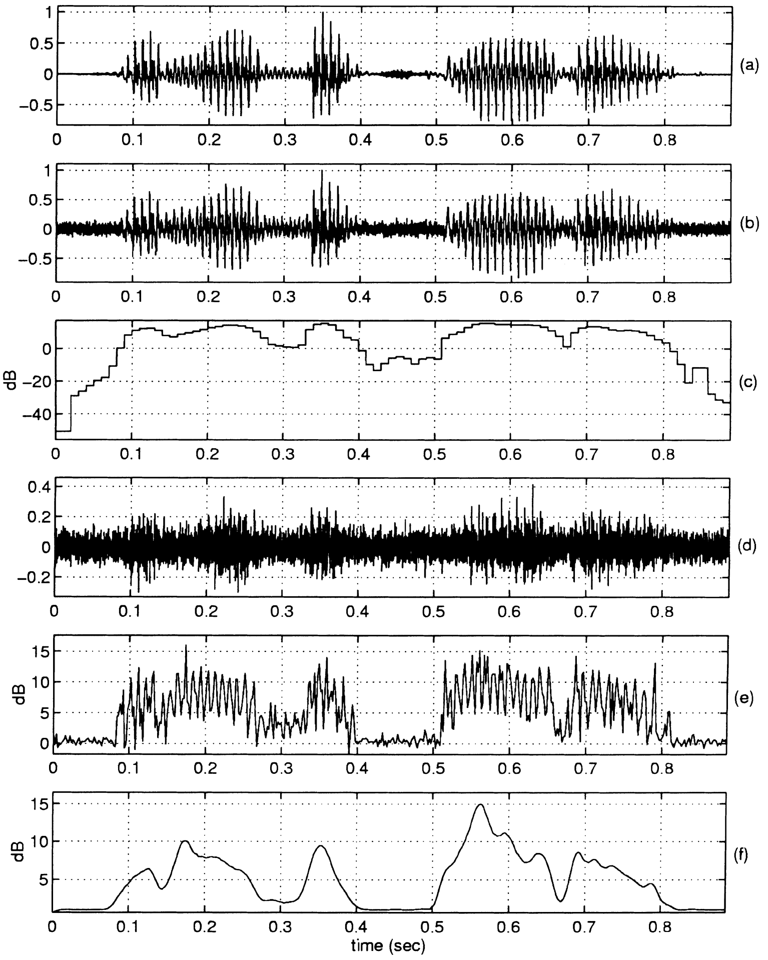

Fig. 1. (a) Speech signal for the utterance "*any dictionary*". (b) Signal with an average SNR of 10 dB. (c) The SNR as a function of time. (d) The 12th order LP residual signal derived from the noisy signal in (b). (e) The ratio of energy values between (d) and (b) for 10 dB SNR case for each 2 ms frame. (f) The ratio curve in (e) smoothed using a 17-point Hamming window.

is the vector of clean signal samples and

$$\boldsymbol{n} = [n_t, n_{t+1}, \ldots, n_{t+N-1}]^{\mathrm{T}}$$

is the vector of noise samples. Let $\boldsymbol{r}$ be the vector of residual error samples derived by inverse filtering the noisy signal $\boldsymbol{x}$ using a pth order LP analysis. The linear prediction coefficients (LPCs) are denoted by $a_0, a_1, a_2, \ldots, a_p$ with $a_0 = 1$. Assuming the initial conditions to be zero, the residual signal vector $\boldsymbol{r}$ may be expressed in matrix form as

$$\boldsymbol{r} = \boldsymbol{A}\boldsymbol{x}, \tag{2}$$

where

$$\boldsymbol{A} = \begin{bmatrix} a_0 & 0 & & \cdots & & 0 \\ a_1 & a_0 & & \cdots & & 0 \\ \vdots & \vdots & \ddots & & & \vdots \\ a_p & a_{p-1} & \cdots & a_0 & \cdots & 0 \\ \vdots & \ddots & & & \ddots & \vdots \\ 0 & \cdots & a_p & \cdots & a_1 & a_0 \end{bmatrix}. \tag{3}$$

An estimate of the clean signal can be obtained by weighting the derived residual error samples appropriately and exciting the LP all-pole filter. The weighted residual error vector $\boldsymbol{r}_{\mathrm{w}}$ can be expressed as

$$\boldsymbol{r}_{\mathrm{w}} = \boldsymbol{W}\boldsymbol{r}, \tag{4}$$

where $\boldsymbol{W} = \mathrm{diag}\,[w(0), w(1), \ldots, w(N-1)]$ is the diagonal $N \times N$ matrix of optimal weights to be estimated. An estimate of the clean signal is given by

$$\hat{\boldsymbol{s}} = \boldsymbol{H}\boldsymbol{r}_{\mathrm{w}}, \tag{5}$$

where

$$\boldsymbol{H} = \boldsymbol{A}^{-1} \tag{6}$$

is the matrix of coefficients of truncated impulse response of the all-pole filter. The truncation effects are assumed to be negligible. Using Eqs. (2) and (4),

$$\hat{\boldsymbol{s}} = \boldsymbol{HWAx} = \boldsymbol{HWAs} + \boldsymbol{HWAn}. \tag{7}$$

The error in reconstruction is given by

$$\boldsymbol{e} = \boldsymbol{s} - \hat{\boldsymbol{s}} = (\boldsymbol{I} - \boldsymbol{HWA})\boldsymbol{s} - \boldsymbol{HWAn}.$$

Using Eq. (6) in the above equation we find

$$\boldsymbol{e} = \boldsymbol{H}(\boldsymbol{I} - \boldsymbol{W})\boldsymbol{As} - \boldsymbol{HWAn}. \tag{8}$$

The energy of the reconstruction error $\boldsymbol{e}$ can be minimized with respect to the weight matrix $\boldsymbol{W}$. But this error criterion does not exploit the masking properties of the human ear (Mermelstein, 1982; Jayant and Noll, 1984; Chen and Gersho, 1995). Hence, a criterion which would be more meaningful perceptually would be the energy of the filtered reconstruction error $\boldsymbol{e}_p$. The filter can be the inverse filter $A(z) = a_0 + a_1 z^{-1} + \cdots + a_p z^{-p}$ of the LP analysis. For an SNR greater than 10 dB it is reasonable to assume that the inverse filter $A(z)$ exhibits valleys at approximately the formant frequencies, although its dynamic range would be low because of noise in the speech signal. Minimization of the energy of the filtered error with respect to $\boldsymbol{W}$ would allow more error in the formant regions and minimizes the error in the valley regions, which is desirable from a perceptual viewpoint. From Eq. (8), the filtered error $\boldsymbol{e}_p$ can be written as

$$\boldsymbol{e}_p = \boldsymbol{Ae} = \boldsymbol{AH}(\boldsymbol{I} - \boldsymbol{W})\boldsymbol{As} - \boldsymbol{AHWAn}. \tag{9}$$

Using Eq. (6) in Eq. (9) we obtain

$$\boldsymbol{e}_p = (\boldsymbol{I} - \boldsymbol{W})\boldsymbol{As} - \boldsymbol{WAn}. \tag{10}$$

Let $\boldsymbol{r}_{\mathrm{s}} = \boldsymbol{As}$ be the signal obtained by filtering the clean signal $\boldsymbol{s}$ using the filter $A(z)$ derived from the noisy signal $\boldsymbol{x}$, and let $\boldsymbol{v} = \boldsymbol{An}$ be the filtered noise in the residual signal domain, then

$$\boldsymbol{e}_p = (\boldsymbol{I} - \boldsymbol{W})\boldsymbol{r}_{\mathrm{s}} - \boldsymbol{Wv}. \tag{11}$$

Assuming that the signal $\boldsymbol{s}$ and noise $\boldsymbol{n}$ are uncorrelated, the cost function

$$\phi(\boldsymbol{W}) = \mathscr{E}\{\|\boldsymbol{e}_p\|^2\} \tag{12}$$

is minimized to obtain the optimum weights as

$$w_{\mathrm{opt}}(k) = \frac{\mathscr{E}\{r_{\mathrm{s}}^2(k)\}}{\mathscr{E}\{r_{\mathrm{s}}^2(k)\} + \mathscr{E}\{v^2(k)\}}, \quad k = 0, 1, \ldots, N-1, \tag{13}$$

where $r_{\mathrm{s}}(k)$ and $v(k)$ are the kth components of $\boldsymbol{r}_{\mathrm{s}}$ and $\boldsymbol{v}$, respectively. If we define the following ratio as an approximate measure of SNR in the residual signal domain:

$$\mathrm{SNR}(k) = \frac{\mathscr{E}\{r_{\mathrm{s}}^2(k)\}}{\mathscr{E}\{v^2(k)\}}, \tag{14}$$

then we have

$$w_{\mathrm{opt}}(k) = \frac{\mathrm{SNR}(k)}{1 + \mathrm{SNR}(k)}. \tag{15}$$

The solution in Eq. (15) is clearly a time domain analogue of the optimal Wiener filter frequency response (Haykin, 1991). Note that in arriving at the result in Eq. (15), no restriction is placed on the noise samples in the vector $\boldsymbol{n}$. The noise samples are only assumed to be uncorrelated with the signal samples in the vector $\boldsymbol{s}$. Since it is difficult to estimate $\mathrm{SNR}(k)$ in practice, $w_{\mathrm{opt}}(k)$ can only be approximated as discussed in Section 3. Note that the optimal weight $w_{\mathrm{opt}}(k)$ in Eq. (15) approaches 1 in the limit when $\mathrm{SNR}(k) \gg 1$ and approaches $\mathrm{SNR}(k)$ itself, when $\mathrm{SNR}(k) \ll 1$. But in our method (presented in Section 3) the weight function used is not exactly same as the optimal weight. Firstly, it is difficult to estimate the $\mathrm{SNR}(k)$ in practice. Secondly, allowing the weight to assume very low values when the $\mathrm{SNR}(k)$ is poor produces distortion in the processed speech. Hence, it is necessary to restrict the minimum value of the weight. Assuming that the noise variance in the residual signal domain is approximately constant, $\mathrm{SNR}(k)$ is proportional to the short-time energy of the residual signal. Hence, the short-time energy values of the residual signal are used to derive the weight function at the finer (1–2 ms) level.

2.3. Nature of LP residual signal

An experiment was conducted to demonstrate the effect of processing the LP residual signal of speech and reconstructing the speech using only a part of the residual signal after the instant of glottal closure. From the clean speech the voiced/unvoiced/silence segments and the instants of significant excitation were identified (Smits and Yegnanarayana, 1995; Yegnanarayana and Teunen, 1994). The LP residual signal of noisy speech was modified retaining only the 2 ms portions of the residual signal around the instants of excitation. The modified residual signal was used to excite the time-varying all-pole filter to regenerate the speech signal. The resulting speech was significantly enhanced without causing serious distortion. This is because the high SNR segments of noisy speech were retained in the reconstructed speech. Note that the all-pole filter derived from the noisy speech may not represent the spectral features of the clean speech accurately. The coefficients of the filter were used mainly to derive the noisy residual signal by inverse filtering. Retaining the waveform bells around the glottal closure produces good quality speech as was demonstrated in PSOLA based Text-to-Speech system (TTS) (Hamon et al., 1989).

The LP residual signal (Fig. 1(d)) may be derived for the noisy speech using a frame of 20 ms duration and a frame rate of about 100 frames per second. Even in the LP residual signal of noisy speech, the SNR is a function of time or frequency. Inverse filtering reduces the correlation between samples existing in the noisy speech signal. Since the residual signal samples are less correlated, the SNR as a function of time can be studied using much smaller windows (1–2 ms) than the windows (10–30 ms) normally used in the short-time spectral analysis. The truncation effects of the analysis window are significantly reduced in the residual signal (Yegnanarayana et al., 1998a). For each small window of the residual signal, the energy ratio of the noisy speech signal and the corresponding portion of the residual signal gives an indication of the amount of reduction in the correlation of the signal samples. This also gives an indication of how much the signal spectrum is flattened in the residual signal. If the signal spectrum is already flat, then the ratio of the energies of the noisy signal and the residual signal in the short (1–2 ms) window will be nearly unity. Otherwise, the ratio will be quite large. Note that for noise-like segments this ratio of the energies will be nearly unity. Thus the ratio of the energies gives an indication of the signal and noise regions of the signal. The ratio of the energy values for a 10 dB SNR situation computed for each 2 ms frame is shown in Fig. 1(e). Note that even weak signal regions are discernible in the ratio plots. The ratio can be interpreted as the inverse of spectral flatness of the noisy signal, the minimum inverse flatness being one, corresponding to the energy ratio of 0 dB.

Since the correlation between the residual signal samples is low, these samples can be manipulated to some extent without producing significant distortion in the reconstructed speech (Yegnanarayana and Satyanarayana Murthy, 1996). It is this manipulative capability of the residual signal we would like to exploit for enhancement of speech.

3. Manipulation of LP residual signal

The basic principle of our approach for speech enhancement is to identify the low SNR regions in the LP residual signal, and derive a weight function for the residual signal which will reduce the energy in the low SNR regions relative to the high SNR regions of the noisy signal. The residual signal samples are multiplied with the weight function. The modified residual signal is used to excite the time-varying all-pole filter to generate the enhanced speech. Speech enhancement is carried out at three levels: (a) at gross level, based on the overall smoothed inverse spectral flatness characteristics; (b) at finer level (1–2 ms), based on the relative energies of the residual signal between adjacent frames; and (c) at spectral level, to enhance the features in the spectrum that could not be affected by the fine level operations.

3.1. Gross temporal level

At the gross level the regions corresponding to low and high SNR regions are identified from the characteristics of the LP residual signal. A weight function for the residual signal samples is derived based on the smoothed inverse spectral flatness characteristics of the noisy speech signal. The spectral flatness characteristics are derived by comparing the energy in the residual signal with the energy in the noisy speech signal in each short interval of about 2 ms.

Inverse filtering the noisy speech signal using the time-varying LP coefficients will give the residual signal. The ratio of the noisy speech signal energy to the residual signal energy in dB for each nonoverlapping frame of 2 ms gives an indication of the inverse spectral flatness as a function of time. The inverse spectral flatness plot is smoothed using a 17-point Hamming window. The smoothed inverse flatness plot shown in Fig. 1(f) clearly indicates the low and high SNR regions. The low SNR (noisy) regions have an inverse flatness close to unity (0 dB), and the high SNR (signal) regions have larger inverse flatness values. Note that for noise-like segments the inverse flatness will be close to unity. Unvoiced segments can be distinguished from noisy segments by the higher residual signal energy value for the unvoiced region compared to the energy value in the (noisy) silence region (see Fig. 1(c)). A weight function is derived from the smoothed inverse flatness characteristics in such a way that the residual signal samples in the regions corresponding to low values of the inverse flatness are reduced relative to the residual signal samples in the regions corresponding to high values of the inverse flatness.

A mapping function of the type shown in Fig. 2 can be used to map the smoothed inverse spectral flatness values to the weight values for each short (2 ms) frame of residual signal. The mapping function is of the type $\tanh(x)$. The purpose of the nonlinear mapping function is to enhance the contrast between the value of the inverse spectral flatness in the speech signal regions and its value in the background noise regions. The weight values for each frame are further smoothed using a 2 ms window to compute the running average across time. Thus we can generate a weight value for each sample of the residual signal as shown in Fig. 3(a).

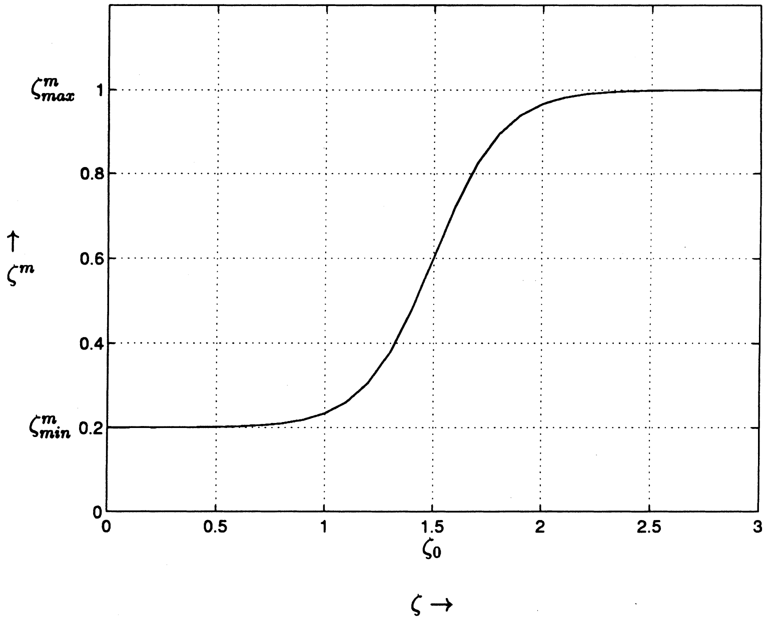

Fig. 2. Mapping function to generate the mapped energy ratio values (ζ^m) from the energy ratio values (ζ). The mapping function $\zeta^m = \frac{1}{2}(\zeta^m_{\max} - \zeta^m_{\min})\ \tanh(\alpha_g \pi(\zeta - \zeta_0)) + \frac{1}{2}(\zeta^m_{\max} + \zeta^m_{\min})$ is shown for $\alpha_g = 0.75$ and $\zeta_0 = 1.50$.

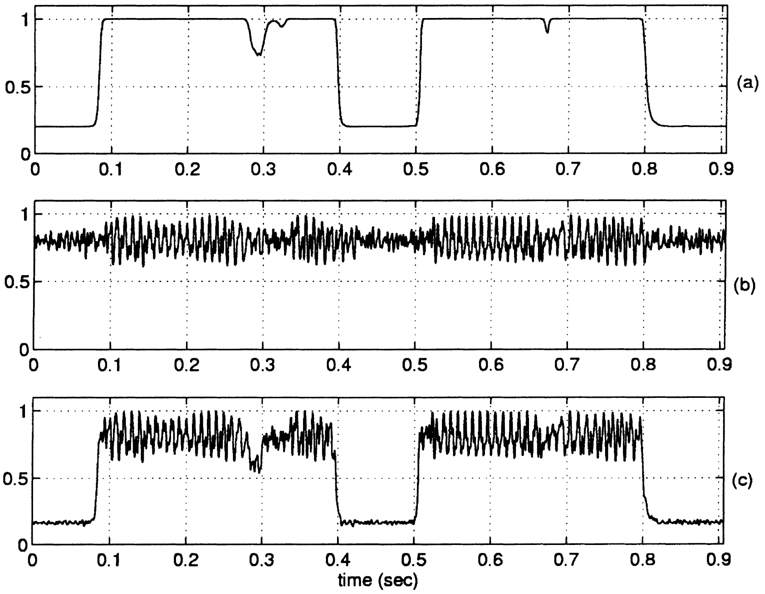

Fig. 3. Weight functions for the LP residual signal. (a) Gross weight function. (b) Fine weight function. (c) Final weight function.

The residual signal samples are multiplied with this weight function to generate a modified residual signal.

The noisy and the enhanced signals along with their spectrograms are shown in Fig. 4. The figure shows the reduction in the energy in the noisy segments relative to the speech segments. On listening, we notice a significant reduction in the annoyance due to the background noise. However, due to sudden change from low noise to the noisy speech regions, the change can be perceived in the enhanced speech. It is possible to trade between the annoyance and speech quality by adjusting the thresholds in the mapping function shown in Fig. 2. The more the reduction in the noise level in the low SNR regions relative to the noise in the high SNR regions, the better will be the speech quality. But then there will be more annoyance due to sudden rise in the background noise. Further improvement can be obtained by manipulating the residual signal at the finer level as discussed in Section 3.2.

3.2. Finer temporal level

From the spectrogram in Fig. 4(f) we notice that the noise in the enhanced speech regions is distributed uniformly across frequency in the spectrum. This causes annoyance due to abrupt change from low noise to high noise regions in the time domain. Also the speech formant features are masked due to noise filling up the low amplitude portions in the frequency domain. Further enhancement at finer levels in the speech segments, especially in the voiced regions, may improve the quality and reduce the annoyance.

For voiced segments, if the SNR is low in some short (1–2 ms) segments, then the residual signal in those regions can be given lower weightage compared to the adjacent higher SNR segments. This is likely to happen for the regions corresponding to the open glottis portion in each glottal cycle due to damping of the formants. The fluctuations in the residual signal energy contour for short (2 ms) segments illustrate the energy differences between adjacent segments. A weight function at the fine level can be derived from the residual signal energy plot to deemphasize the segments corresponding to the valleys relative to the segments corresponding to the peaks. But for noisy speech, the residual signal is noisy and so the energy of the short segment of the residual signal may not be reliable for deriving the weight. Hence, the Frobenius norm

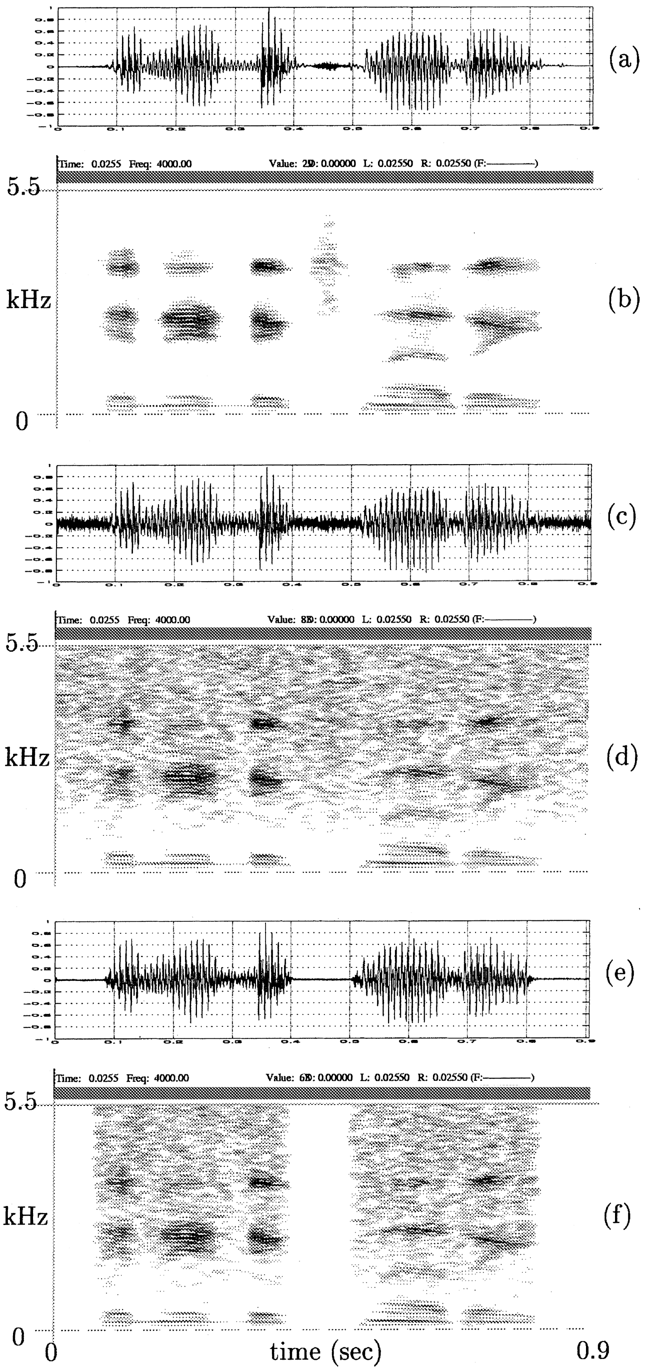

Fig. 4. Results of enhancement of speech degraded by additive white noise. (a) Clean speech. (c) Speech signal at 10 dB SNR. (e) Enhanced speech signal obtained using gross level weighting. (b), (d), (f) Spectrograms for the signals in (a), (c), (e), respectively.

(Leon, 1990) of the Toeplitz prediction matrix (see Eq. (16)) constructed using the noisy speech samples in a frame of 2 ms duration is used to represent the short-time energy of the corresponding frame of LP residual signal (Satyanarayana Murthy and Yegnanarayana, 1998). This approach has the advantage of exploiting the envelope information in the noisy speech waveform. The Toeplitz prediction matrix $\boldsymbol{X}$ is given by

$$\boldsymbol{X} = \begin{bmatrix} x_{p+1} & x_p & \cdots & x_1 \\ x_{p+2} & x_{p+1} & \cdots & x_2 \\ & & \ddots & \vdots \\ \vdots & \vdots & & x_{p+1} \\ & & & \vdots \\ x_M & x_{M-1} & \cdots & x_{M-p} \end{bmatrix}, \qquad (16)$$

where $x_1, x_2, \ldots, x_M$ are the noisy speech samples in a frame of length M samples, which is 16 for 2 ms duration at 8 kHz sampling. The LP order p is taken as 10. The Frobenius norm is computed for every sample. The weight function is derived using the logarithm of the ratio of the Frobenius norm of the present frame to the Frobenius norm of the frame 2 ms prior to the present frame. A mapping function of the type shown in Fig. 2 is used to map the log ratio values to the weight values for each sample of the signal. The objective of the mapping function is to control the relative emphasis of high SNR segments over low SNR segments in short (2 ms) intervals. The maximum change is restricted to the interval 0.2–1.0. The finer weight function is shown in Fig. 3(b). The overall weight function is obtained by multiplying the gross weight function derived from the smoothed inverse flatness plot with the fine weight function. The final weight function for the residual signal samples is shown in Fig. 3(c). Enhanced speech is generated by exciting the time-varying all-pole filter with this weighted residual signal. Spectrograms of the enhanced speech along with the spectrograms for noisy speech and the speech enhanced using only gross level weighting of the residual signal are shown in Fig. 5. From the spectrogram in Fig. 5(c) we observe that the spectrum of the signal is significantly enhanced in the voiced regions. The quality of speech is significantly better than in the case of the gross level modification.

3.3. Spectral level

In the reconstruction of enhanced speech, even though the LP residual signal is deemphasized in the low SNR regions, the all-pole filters derived from the 20 ms segments dominate the system characteristics in the reconstructed speech signal. To improve the system characteristics at the spectral level, the LPCs for shorter (1–3 ms) segments need to be obtained from noisy speech. This will make the all-pole filter for the high SNR segments closer to the true one. For other segments the amplitude of the output signal is reduced in the reconstruction due to deemphasis of the corresponding residual signal. But unfortunately, we do not have a good method of estimating the all-pole filter for short (1–3 ms) segments.

One way to achieve spectral manipulation indirectly is to perform a low order LP analysis on the preemphasized speech signal. A 7th order LP analysis is performed using 5 ms Hamming windowed segments overlapped by 2 ms. Due to the Hamming window, the effective duration of the signal used for analysis is less than 5 ms. The residual signal is computed by passing the speech signal through the inverse filter. The residual signal is manipulated as described before. The modified residual signal is used to excite the time–varying all-pole filter, updated every 2 ms, to generate the enhanced speech. The different steps in the algorithm are presented in Table 1.

4. Experimental results

Examples are given in this section to demonstrate the performance of the proposed method for different types of noises. The degradation is gradual and graceful as the noise level is increased. This is because the LP analysis tends to be less accurate as the SNR reduces. It is important to note that the thresholds for deriving the weight function could be adjusted so as to obtain an acceptable trade-off between reduction in annoyance due to

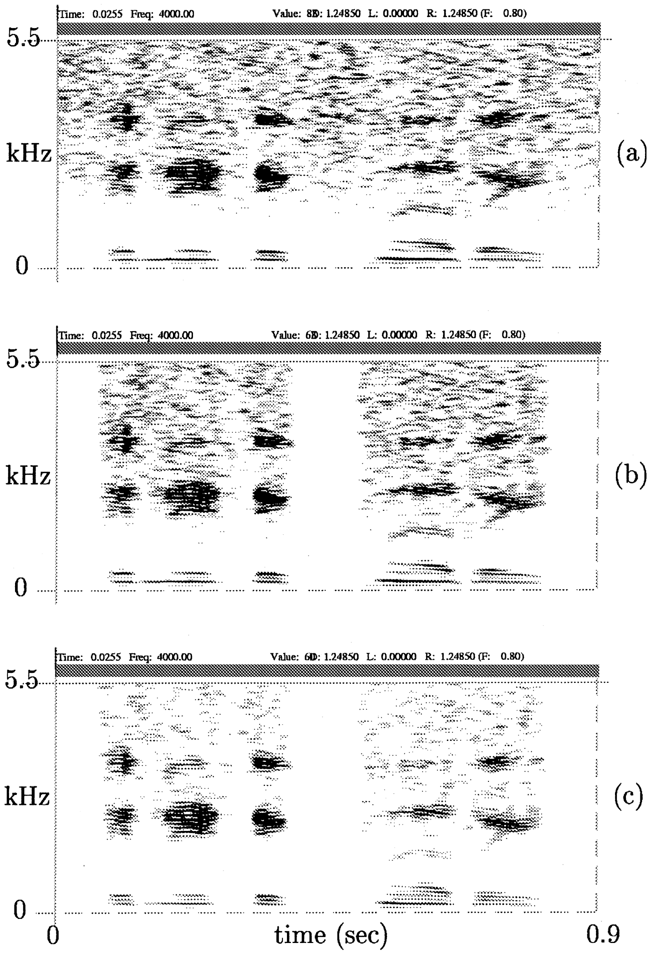

Fig. 5. (a) Spectrogram for 10 dB SNR speech. (b) Spectrogram for enhanced speech using gross level weighting of the residual signal. (c) Spectrogram for enhanced speech using gross and fine level weighting of the residual signal.

noise and degradation in speech quality, based on perceptual impression of the enhanced speech. However, once the listener sets the thresholds to suit his preference, they need not be adjusted again.

4.1. Studies on different types of noises

The proposed method for speech enhancement works well even for colored additive noise. Fig. 6(a) shows the spectrogram of speech corrupted by noise recorded in the cockpit of an F16 aircraft (Website, 1997). The average SNR is adjusted to 10 dB. We notice from the spectrograms that the cockpit noise exhibits both broadband as well as narrowband (spectral lines at approximately 3000 and 4500 Hz) characteristics. Fig. 6(b) shows the spectrogram of enhanced speech. The enhancement was carried out using the algorithm proposed in the previous section in three iterations. We found that the enhancement was better when carried out in smaller steps over two or three iterations rather than in one step. In each iteration, mild enhancement can be obtained by using suitable values for the thresholds used for the mapping function in Fig. 2. The thresholds were chosen so as to achieve mild enhancement in each iteration and are kept constant in all the iterations.

Table 1
Algorithm for processing noisy speech for enhancement

Computation of the gross weight function

- Calculate the linear prediction (LP) residual signal using a speech frame of size 20 ms, overlapping by 10 ms, Hamming window and a 10th order LP analysis by autocorrelation method. The analysis is performed on the preemphasized speech signal.
- Calculate the ratio of the noisy speech signal energy and the LP residual signal energy for each nonoverlapping 2 ms frame. The ratio gives the inverse spectral flatness value for each 2 ms frame.
- Smooth the inverse spectral flatness curve using a 17-point Hamming window. The smoothed spectral flatness value is denoted by ζ_k for the kth frame.
- Obtain the output ζ_k^m of the mapping function

$$\zeta_k^m = \left(\frac{\zeta_{\max}^m - \zeta_{\min}^m}{2}\right) \tanh\left(\alpha_g\ \pi\ (\zeta_k - \zeta_0)\right) + \left(\frac{\zeta_{\max}^m + \zeta_{\min}^m}{2}\right)$$

  from ζ_k (see Fig. 2).
- Obtain the gross weight function by repeating each mapped value ζ_k^m 16 times (2 ms at 8 kHz sampling) and smoothing it with a 2 ms mean smoothing filter. This generates a gross weight value w_n^{gross} for every sampling instant n.

Computation of the fine weight function

- Compute the Frobenius norm of the Toeplitz prediction matrix constructed using the noisy speech samples in each 2 ms frame, for every sampling instant n.
- Compute the logarithm of the ratio of Frobenius norms of the current frame at the nth sampling instant to the Frobenius norm of the frame 2 ms (= 16 sampling instants) prior to the current frame. Normalize the log ratio w.r.t. the maximum value. Obtain the fine weight function w_n^{fine} by mapping the normalized log ratio using the function

$$w_n^{\text{fine}} = \left(\frac{w_{\max}^{\text{fine}} - w_{\min}^{\text{fine}}}{2}\right) \tanh\left(\alpha_f\ \pi\ y_n\right) + \left(\frac{w_{\max}^{\text{fine}} + w_{\min}^{\text{fine}}}{2}\right)$$

  which is similar to the function shown in Fig. 2. w_n^{fine} is the fine weight value at the nth sampling instant, y_n is the normalised log ratio of Frobenius norms at n, $w_{\max}^{\text{fine}}$ (= 1) is the maximum mapped value, $w_{\min}^{\text{fine}}$ (= 0.6) is the minimum mapped value and α_f (= 0.75) is a positive constant.

Linear prediction analysis

- Calculate the linear prediction (LP) residual signal using a speech frame of size 5 ms, overlapping by 2 ms, Hamming window and a 7th order LP analysis by autocorrelation method. The analysis is performed on the preemphasised speech signal.

Synthesis of enhanced speech

- Multiply the two weight functions w_n^{gross} and w_n^{fine} to generate the overall weight function.
- Multiply the LP residual signal obtained above using 5 ms segments of the speech signal by the overall weight function. The weighted residual signal is used to excite the time-varying all-pole filter updated every 2 ms, to generate enhanced speech.

The method was tested for real signals where the speech signal and noise were recorded simultaneously. Fig. 7(a), (c) and (e) show the clean, degraded and processed speech signals, respectively. The clean speech and degraded speech were collected simultaneously by two microphones, one placed at a distance of 0.1 m from the speaker and the other placed 1.2 m away. The speech signal shown in Fig. 7(a) corresponds to the sentence "*She had your dark suit in greasy wash water all year*" spoken by a male speaker and is taken from the TIMIT database. It can be seen in Fig. 7(c) that the degraded speech has small amount of room reverberation in addition to ambient (air-conditioner) noise. The ambient noise has lowpass spectral characteristics and some narrowband spectral components. In fact there is ambient noise present even in the clean speech signal in Fig. 7(a). The speech signal in Fig. 7(c) was preemphasized before processing. The speech signal processed using the proposed algorithm and its spectrogram are shown in Fig. 7(e) and (f), respectively. It can be seen from the Fig. 7(e) and (f) that the noise level is significantly attenuated, especially in the silence regions. It is important to note that the gross weight function provides mild attenuation of the reverberation tails. Informal listening confirms that there is reduction of the annoyance due to noise in the processed speech signal, without introducing significant distortion.

The proposed method was also tested on female speech. The experimental setup for data collection

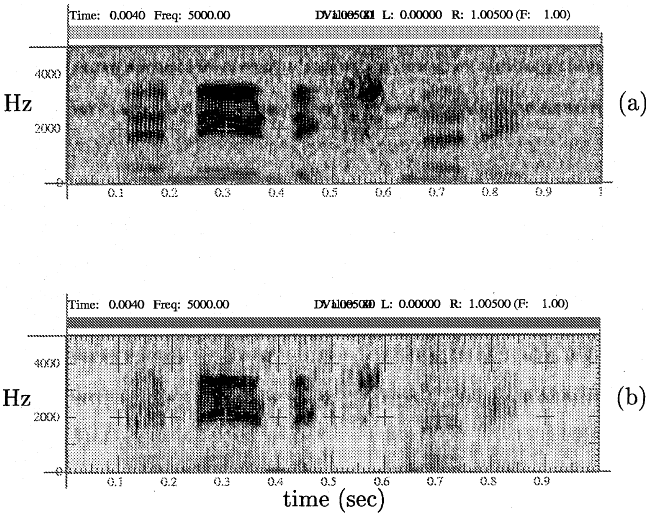

Fig. 6. (a) Spectrogram for 10 dB SNR speech. The speech is corrupted by aircraft cockpit noise. (b) Spectrogram for enhanced speech using spectral level manipulation besides gross and fine level weighting of the LP residual signal. The speech is enhanced using three iterations.

was the same as that used for collection of male speech mentioned above. Fig. 8(a) shows the clean speech signal corresponding to the sentence "*She had your dark suit in greasy wash water all year*" taken from the TIMIT database. The spectrograms of clean, degraded and processed speech signals are shown in Fig. 8(b), (d) and (f), respectively. The improvement obtained due to processing can be clearly seen in the spectrogram in Fig. 8(f). Informal listening confirms the improvement obtained due to processing. It is important to note that the same thresholds were used for the mapping functions in all the experiments.

4.2. *Performance of the method for different parameter settings*

A comparison of the performance of the proposed method for two different settings of the parameters of the mapping function is shown in Fig. 9 for the case of female speech. A comparison with the performance of the spectral subtraction method (Boll, 1979) is also given in the same figure (Fig. 9(c)). The speech signal used for this comparison is the same as the one shown in Fig. 8(c). Fig. 9(a) and (b) show the spectrograms of the processed speech signals for the parameter settings A and B, respectively, given in Table 2. The parameter settings for case A are chosen such that a mild enhancement of the noisy speech signal is obtained without introducing distortion in the processed signal. The emphasis of speech regions with respect to the background noise regions is relatively more for the parameter settings for case B compared to that for case A. But in this case mild distortion is perceived in the processed signal.

Although the spectrogram in Fig. 9(a) does not show significant improvement when compared to the spectrogram of noisy speech in Fig. 8(d), the improvement can be clearly perceived while listening. The spectrogram in Fig. 9(b) shows a significant improvement when compared with the spectrogram of the noisy speech in Fig. 8(d). Note that the weighting of the residual signal at the fine level (i.e. relative emphasis of the residual signal samples within a glottal cycle) should be mild to avoid distortion in the processed speech. In the voiced regions the spectrogram in Fig. 9(b) appears cleaner compared to the spectrogram in Fig. 9(a). For the case of speech processed using the spectral subtraction method we observe that weak spectral peaks appear randomly in the spectrogram in Fig. 9(c). These random spectral peaks give rise to musical noise.

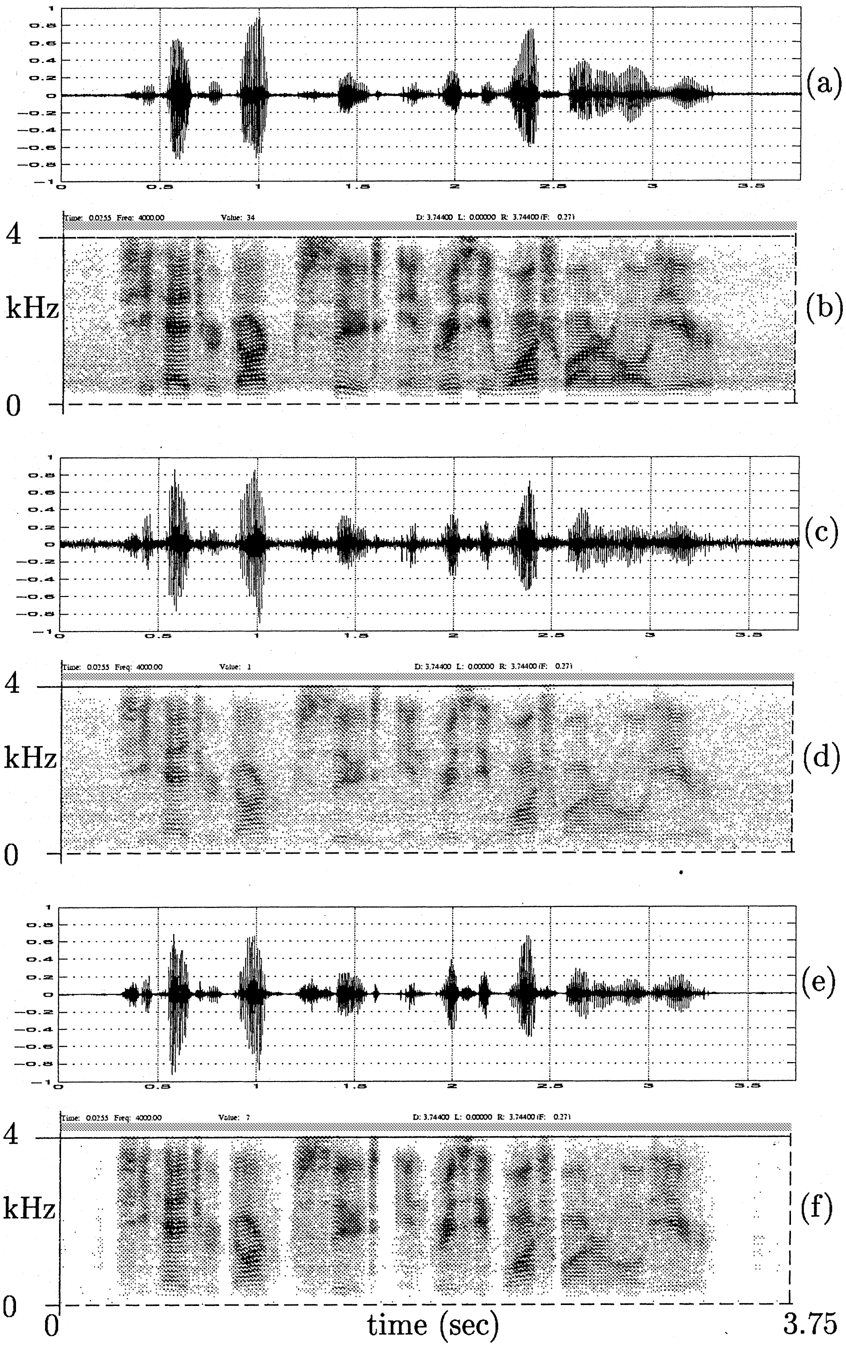

Fig. 7. Results of enhancement of male speech degraded by ambient noise. (a) Clean speech. (b) Spectrogram of clean speech. (c) Speech degraded by noise. (d) Spectrogram of speech degraded by noise. (e) Speech processed using the proposed algorithm. (f) Spectrogram of processed speech.

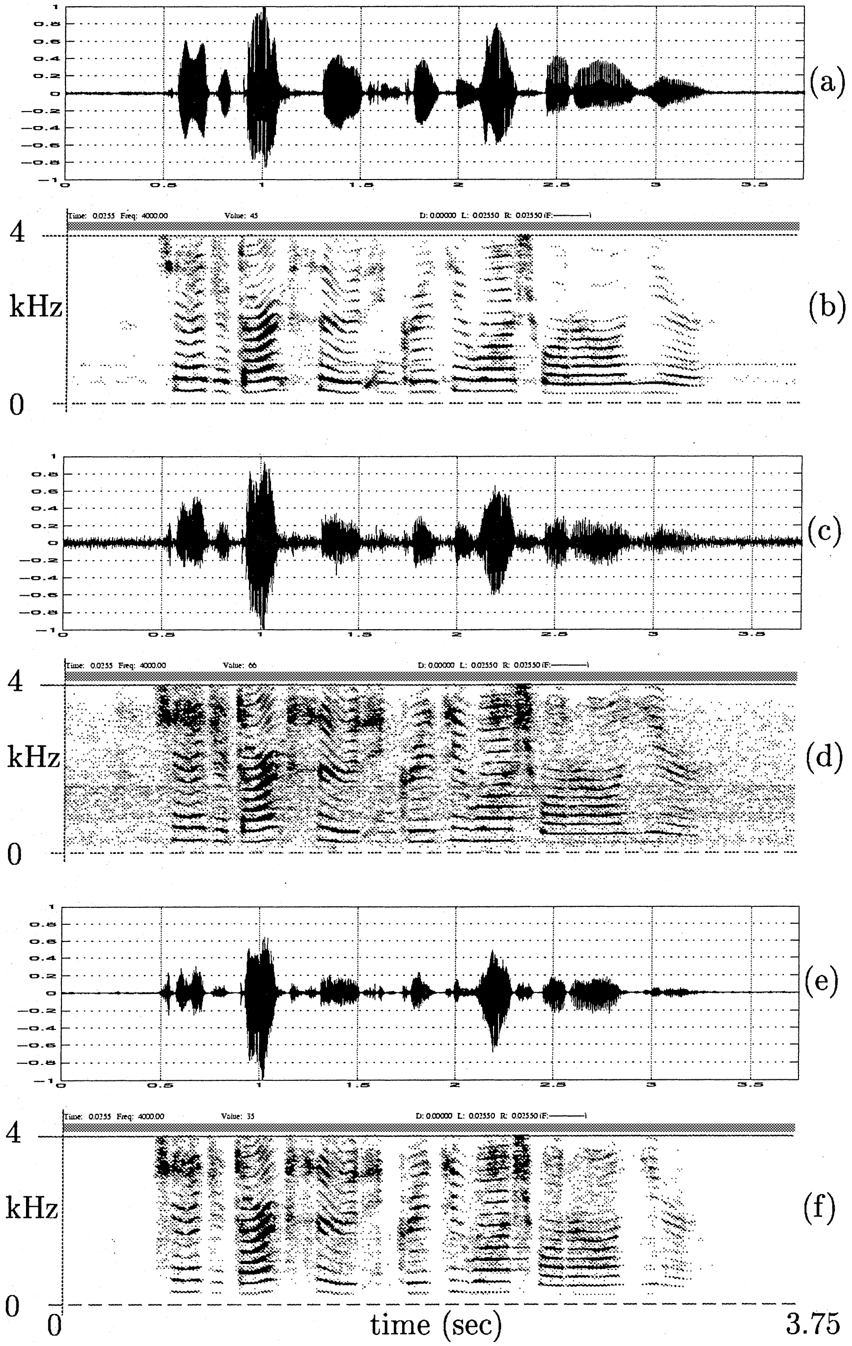

Fig. 8. Results of enhancement of female speech degraded by ambient noise. (a) Clean speech. (b) Spectrogram of clean speech. (c) Speech degraded by noise. (d) Spectrogram of speech degraded by noise. (e) Speech processed using the proposed algorithm. (f) Spectrogram of processed speech.

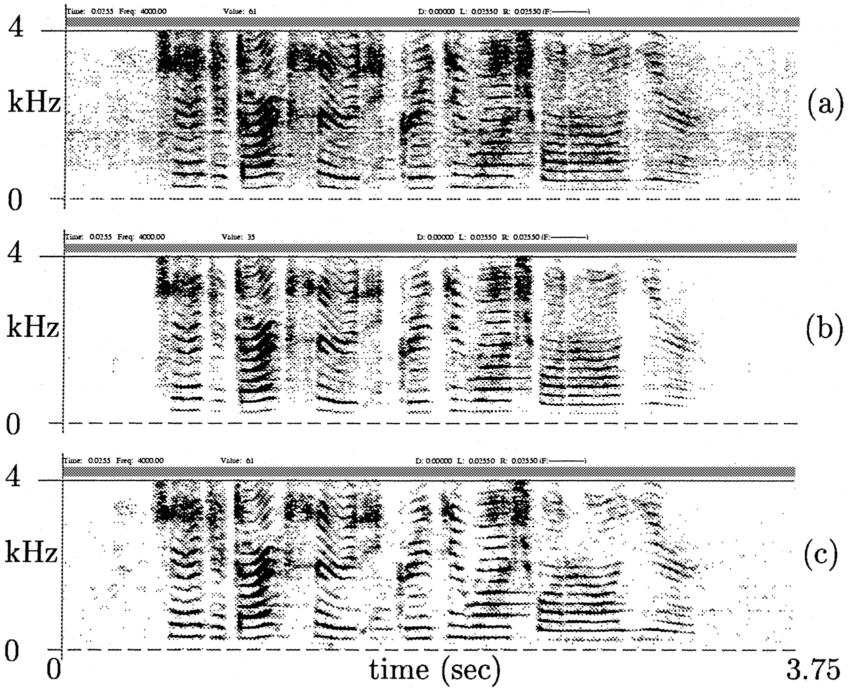

Fig. 9. Comparison of results of enhancement of the proposed method with spectral subtraction for female speech degraded by ambient noise. Spectrograms of speech processed using the proposed algorithm for the parameter settings of (a) case A and (b) case B in Table 2. (c) Spectrogram of speech processed using the spectral subtraction algorithm.

Table 2
Two different settings of the parameters for the mapping functions

	$\zeta^{m}_{\max}$	$\zeta^{m}_{\min}$	α_g	ζ_0	$w^{\text{fine}}_{\max}$	$w^{\text{fine}}_{\min}$	α_f
Case A	1.0	0.1	1.0	1.5	1.0	0.25	0.75
Case B	1.0	0.05	2.0	2.0	1.0	0.6	0.75

5. Summary and conclusions

In this paper we have presented a new approach for enhancement of speech based on LP residual signal. The method uses the fact that in noisy speech the SNR is a function of time and frequency. By enhancing the high SNR regions relative to the low SNR regions, the annoyance due to background noise is reduced without significantly distorting the speech. This is accomplished by identifying the low and high SNR regions based on the characteristics of the spectral flatness in short (2 ms) time frames. The spectral flatness information is derived using the ratio of energies in the LP residual signal of the speech and the noisy signal. Inverse spectral flatness characteristics are used to derive a weight function for the residual signal at gross level, and the Frobenius norm of short (2 ms) segments of the speech signal is used to derive the weight function at finer level. The two weight functions are multiplied to get the overall weight function for the residual signal. The method works since the residual signal samples are nearly uncorrelated, and hence can be manipulated without significantly affecting the quality of the speech regenerated from the modified residual signal. Since no direct manipulation in different frequency bands is involved, this method does not produce the type of distortion which the spectral subtraction and parameter smoothing methods produce.

The objective in this study is to enhance speech over background noise, and not noise suppression or elimination. In fact even a small (3–6 dB) improvement in SNR of noisy speech may give relief to the listener. This study suggests that speech enhancement methods must aim to bring down the annoyance due to noise by mild enhancements.

The setting of various thresholds in the processing is primarily dictated by the listener's tolerance to annoyance due to noise and preference to speech quality. The various parameter values used in the processing, such as LP order, analysis frame size, thresholds of the mapping function etc., are not critical. The choice of the parameters depends on listener's preference, as the effect of these parameters on the resulting quality of the enhanced speech is gradual and not abrupt. Another important feature is that the method does not depend on the pitch of the voice. There is no direct manipulation of the spectrum. However, a better estimation of the vocal tract system characteristics is needed to improve the enhancement at the spectral level.

The proposed method reduces the annoyance due to additive noise but is not very useful in reducing the annoyance due to reverberation. But an approach based on emphasizing the high signal energy regions relative to the low signal energy regions can be developed for enhancement of reverberant speech also (Yegnanarayana et al., 1998b).

In our opinion the proposed approach is different from many of the methods available for processing degraded speech. There is scope for significant improvement by studying the effects of various parameters on the perceptual quality of the enhanced speech. Moreover, this approach may be combined with well known spectrum-based methods for speech enhancement to obtain a better quality of enhanced speech for various types of degradations.

References

Ananthapadmanabha, T.V., Yegnanarayana, B., 1979. Epoch extraction from linear prediction residual for identification of closed glottis interval. IEEE Trans. Acoust. Speech Signal Processing ASSP-27, 309–319.

Avendaño, C., Hermansky, H., Vis, M., Bayya, A., 1996. Adaptive speech enhancement using frequency-specific SNR estimates. In: Proceeding of III IEEE Workshop on Interactive Voice Technology for Telecommunications Applications. Basking Ridge, New Jersey, pp. 65–68.

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

Chen, J.H., Gersho, A., 1995. Adaptive postfiltering for quality enhancement of coded speech. IEEE Trans. Speech Audio Processing 3, 59–71.

Cheng, Y.M., O'Shaughnessy, D., 1991. Speech enhancement based conceptually on auditory evidence. IEEE Trans. Signal Processing 39, 1943–1954.

Cooper, F.S., 1980. Acoustics in human communication: Evolving ideas about the nature of speech. J. Acoust. Soc. Amer. 68, 18–21.

Deller, J.R., Proakis, J.G., Hansen, J.H.L., 1993. Discrete-Time Processsing of Speech Signals. Macmillan, New York.

Ephraim, Y., 1992. Statistical model based speech enhancement systems. Proc. IEEE 80, 1526–1555.

Ephraim, Y., Van Trees, H.L., 1995. A signal subspace approach for speech enhancement. IEEE Trans. Speech Audio Processing 3, 251–266.

Erell, A., Weintraub, M., 1994. Estimation of noise-corrupted speech DFT-spectrum using the pitch period. IEEE Trans. Speech Audio Processing 2, 1–8.

Fant, G., 1993. Some problems in voice source analysis. Speech Communication 13 (1–2), 7–22.

Gibson, J.D., Koo, B., Gray, S.D., 1991. Filtering of colored noise for speech enhancement and coding. IEEE Trans. Signal Processing 39, 1732–1742.

Hamon, C., Moulines, E., Charpentier, F.J., 1989. A diphone synthesis system based on time domain prosodic modifications of speech. In: Proceedings of IEEE International Conference Acoustic Speech and Signal Processing. Glasgow, Scotland, pp. 238–241.

Haykin, S., 1991. Adaptive Filter Theory. Prentice-Hall, Englewood Cliffs, NJ.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. Speech Audio Processing 2, 578–589.

Jayant, N.S., Noll, P., 1984. Digital Coding of Waveforms – Principles and Applications. Prentice Hall, Englewood Cliffs, NJ.

Junqua, J.C., Haton, J.P., 1996. Robustness in Automatic Speech Recognition: Fundamentals and Applications. Kluwer Academic Publishers, Boston.

Le Bouquin, R., 1996. Enhancement of noisy speech signals: Application to mobile radio communications. Speech Communication 18 (1), 3–19.

Lee, K.Y., Shirai, K., 1996. Efficient recursive estimation for speech enhancement in colored noise. IEEE Signal Processing Lett. 3, 196–199.

Leon, S.J., 1990. Linear Algebra with Applications. Macmillan, New York.

Lim, J.S., Oppenheim, A.V., 1978. All-pole modeling of degraded speech. IEEE Trans. Acoust. Speech Signal Processing ASSP-26, 197–210.

Mermelstein, P., 1982. Threshold of degradation for frequency-distributed band-limited noise in continuous speech. IEEE Trans. Acoust. Speech Signal Processing 72, 1368–1373.

Rose, R.C., Hofstetter, E.M., Reynolds, D.A., 1994. Integrated models of signal and background with application to speaker identification in noise. IEEE Trans. Speech Audio Processing 2, 245–257.

Satyanarayana Murthy, P., Yegnanarayana, B., 1998. Robustness of group-delay-based method for extraction of significant instants of excitation from speech signals. IEEE Trans. Speech Audio Processing, in press.

Smits, R., Yegnanarayana, B., 1995. Determination of instants of significant excitation in speech using group delay functions. IEEE Trans. Speech Audio Processing 3, 325–333.

Website, 1997. http://spib.rice.edu/spib/select_noise.html, IEEE Signal Processing Information Base.

Yegnanarayana, B., Ramachandran, V.R., 1992. Group delay processing of speech signals. In: Proceeding of ESCA Workshop on Comparing Speech Signal Representation. Sheffield, England, pp. 411–418.

Yegnanarayana, B., Satyanarayana Murthy, P., 1996. Source-system windowing for speech analysis and synthesis. IEEE Trans. Speech Audio Processing 4, 133–137.

Yegnanarayana, B., Teunen, R., 1994. Prosodic manipulation of speech using knowledge of instants of significant excitation, Tech. Rep. 1029, Institute for Perception Research, Eindhoven, The Netherlands.

Yegnanarayana, B., d'Allesandro, C., Darsinos, V., 1998a. An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. IEEE Trans. Speech Audio Processing 6, 1–11.

Yegnanarayana, B., Satyanarayana Murthy, P., Avendaño, C., Hermansky, H., 1998b. Enhancement of reverberant speech using LP residual. In: Proceedings of IEEE International Conference Acoustic, Speech, and Signal Processing, Vol. 1, Seattle, Washington, pp. 405–408.