Enhancement of Reverberant Speech Using LP Residual Signal

B. Yegnanarayana, *Senior Member, IEEE,* and P. Satyanarayana Murthy

***Abstract*—In this paper, we propose a new method of processing speech degraded by reverberation. The method is based on analysis of short (2 ms) segments of data to enhance the regions in the speech signal having high signal-to-reverberant component ratio (SRR). The short segment analysis shows that SRR is different in different segments of speech. The processing method involves identifying and manipulating the linear prediction residual signal in three different regions of the speech signal, namely, high SRR region, low SRR region, and only reverberation component region. A weight function is derived to modify the linear prediction residual signal. The weighted residual signal samples are used to excite a time-varying all-pole filter to obtain perceptually enhanced speech. The method is robust to noise present in the recorded speech signal. The performance is illustrated through spectrograms, subjective and objective evaluations.**

***Index Terms*—Glottal pulse, linear prediction residual, reverberant speech, short segment analysis, signal-to-reverberant speech, speech enhancement.**

I. Introduction

DEGRADATIONS in speech are caused by additive noise and reverberation. In this paper, we consider enhancement of speech under reverberant conditions. The focus is on the degradation of speech caused in speakerphone-like situation. Speech from a speakerphone contains both the direct component and the reverberant component. The objective of processing is to enhance the signal in the direct component, wherever possible, so that the resulting processed speech is perceived as less reverberant and thus increasing the comfort level for listening.

Normally, degraded (additive or reverberant) speech is processed assuming that the degradation has long term stationary characteristics relative to speech. For example, for additive noise degradation, the noise statistics are estimated from the degraded speech and the long (100–300 ms) term noise effects are subtracted from the short (10–30 ms) time speech spectra [1], [2] to reduce the effects of noise. Due to sharp changes in the subtracted spectra within a frame and across the frames, the resulting processed speech produces significant audible distortions. Thus noise reduction is accomplished at the cost of quality. Likewise, for reverberant speech, the reverberation effects are captured by estimating the impulse response of the room environment from long (500–1000 ms) segments of speech [3]–[5]. The room impulse response is usually long, of the order of 200–300 ms. The reverberant speech is passed through an inverse filter for the room response to dereverberate speech. Here again the estimated long term characteristics are used to filter out its effects from the short (10–30 ms) quasi-stationary segments of speech. The main problems in these approaches for processing degraded speech is that the estimates of the characteristics of the degradations are not good enough to remove their effects in short segments of speech. This is because the level of degradation in terms of signal-to-noise ratio (SNR) is different for different segments of speech. Moreover, the emphasis in many of these approaches seems to be on the degradation and not on speech. In other words, enhancement is sought to be accomplished by suppressing noise from noisy speech.

In noise suppression and dereverberation, there is more emphasis on improving the overall SNR of the degraded speech. In this process most of the attention is given to improve the low SNR regions of speech. When attempting to reduce the degradation in these regions, the natural characteristics of speech are changed, causing significant distortions. This is because all segments of degraded speech are treated equally. In order to improve the overall SNR, it is necessary to reduce the noise in the low SNR regions, which does not produce significant enhancement perceptually.

Methods focusing on characteristics of speech also have been proposed for enhancement of degraded speech [6]–[11]. Some of these methods are based on exploiting the pitch periodicity and high signal energy characteristics in short (10–30 ms) segments of speech [6]–[9], [11], [12]. These methods are mainly applicable for additive noise, and also they depend critically on the periodicity property. Methods for enhancement of reverberant speech generally rely on estimating the impulse response of the inverse system for dereverberation [5]. It is not possible to estimate this response accurately from speech in most situations. In some methods, the room response is collected separately to design the inverse system [13]. The recovery of the average envelope modulation spectrum of the original (anechoic) speech by filtering the time trajectories of spectral bands of reverberant speech has also been proposed [14]–[16]. Several multimicrophone methods have been proposed [17]–[19] for enhancement of speech degraded by room reverberation. The microphone array based methods attempt to enhance the signal in a particular direction and suppress signals arriving from other directions.

Manuscript received April 14, 1998; revised June 22, 1999. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kuldip K. Paliwal.

B. Yegnanarayana is with the Department of Computer Science and Engineering, Indian Institute of Technology, Madras 600 036, India (e-mail: yegna@iitm.ernet.in).

P. S. Murthy is with the Department of Electrical Engineering, Indian Institute of Technology, Madras 600 036, India.

Publisher Item Identifier S 1063-6676(00)03449-0.

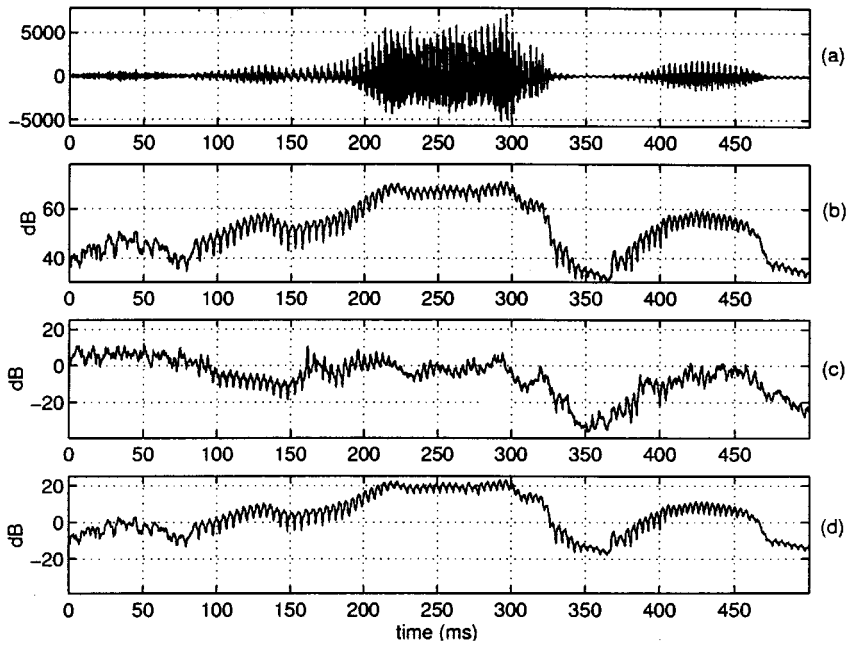


Fig. 1. Variation of short-time signal-to-reverberant sound ratio and signal-to-noise ratio with time for degraded speech: (a) clean speech signal, (b) short-time energy of clean speech computed using 2 ms frames, (c) short-time signal-to-reverberant sound ratio, and (d) short-time signal-to-noise ratio for an average SNR of 10 dB.

There appears to be a need to look at the problem of enhancement of reverberant speech with more emphasis on the direct component of speech at the receiving microphone. In processing, it is necessary to increase the contribution of the direct component relative to the reverberant component [20]. In such an attempt, there will be more emphasis on the speech than on the degradation during the enhancement. This point of view is also reasonable, since speech is a nonstationary signal, with signal energy varying over a wide (about 60 dB) dynamic range both in temporal and spectral domains. Therefore the signal-to-degradation ratio will be varying even within 10–30 ms segments of data. For short (10–30 ms) segments it is difficult to estimate the reverberant component. Moreover, the reverberant component itself will be different in different segments due to its dependence on the energy in the preceding segments of speech. That is, the reverberant component is signal dependent.

It is also essential that we specify our goal in the enhancement of degraded speech. Obviously, complete dereverberation is not a realizable task. Therefore, the emphasis should be on enhancement, but not necessarily enhancement of all segments of speech. There are segments of speech where reverberant component dominates over the direct component. For such segments, there is no point in attempting to enhance the speech part. On the other hand, if regions, where the direct speech signal component is significantly higher compared to the reverberant component, could be identified, then by enhancing speech in such regions the annoyance due to reverberation could be reduced in some segments at least. Likewise, the levels of the signal in the regions with higher reverberation could be reduced, if such regions could be identified. In the regions where there is only a reverberant component, such as silence regions, the levels could be reduced to very low values. Perception of the overall speech is influenced significantly by the high signal energy regions, thus giving an impression of enhancement of degraded speech. Therefore the criterion for improvement need not be based on giving equal emphasis to all the speech segments. It is better to focus on the regions having high direct path signal component.

In this paper, we show that using short segment analysis it is indeed possible to locate the segments in the degraded speech where the direct component is higher than the reverberant component. These segments are usually much shorter than the glottal cycle. The proposed approach is different from the existing methods, as there is more emphasis on the characteristics of speech, and also the analysis segments are much shorter (1–3 ms) compared to the normal frame size (10–30 ms) used in speech analysis. In Section II, we discuss the model of reverberant speech and some of its characteristics. By studying the effects of degradation in short (1–3 ms) segments, we obtain clues that can be used for processing the reverberant speech. In Section III, steps for processing degraded speech are discussed. In particular, the importance of processing the linear prediction (LP) residual signal is emphasized. We present some experimental results in Section IV. The improvement in the processed speech is demonstrated through the signal waveform, short-time spectra, and spectrograms.

II. Characteristics of Reverberant Speech

In this section, we will examine the characteristics of reverberant speech to determine clues for processing speech for enhancement. Throughout the discussion we will examine the similarities and differences in the characteristics of reverberant speech and speech corrupted by additive noise. For this purpose, we consider the following models for reverberant speech and noisy speech.

$$\text{Reverberant speech:} \quad x(n) = s(n) + \sum_{k=1}^{N} b_k s(n - n_k) \quad (1)$$

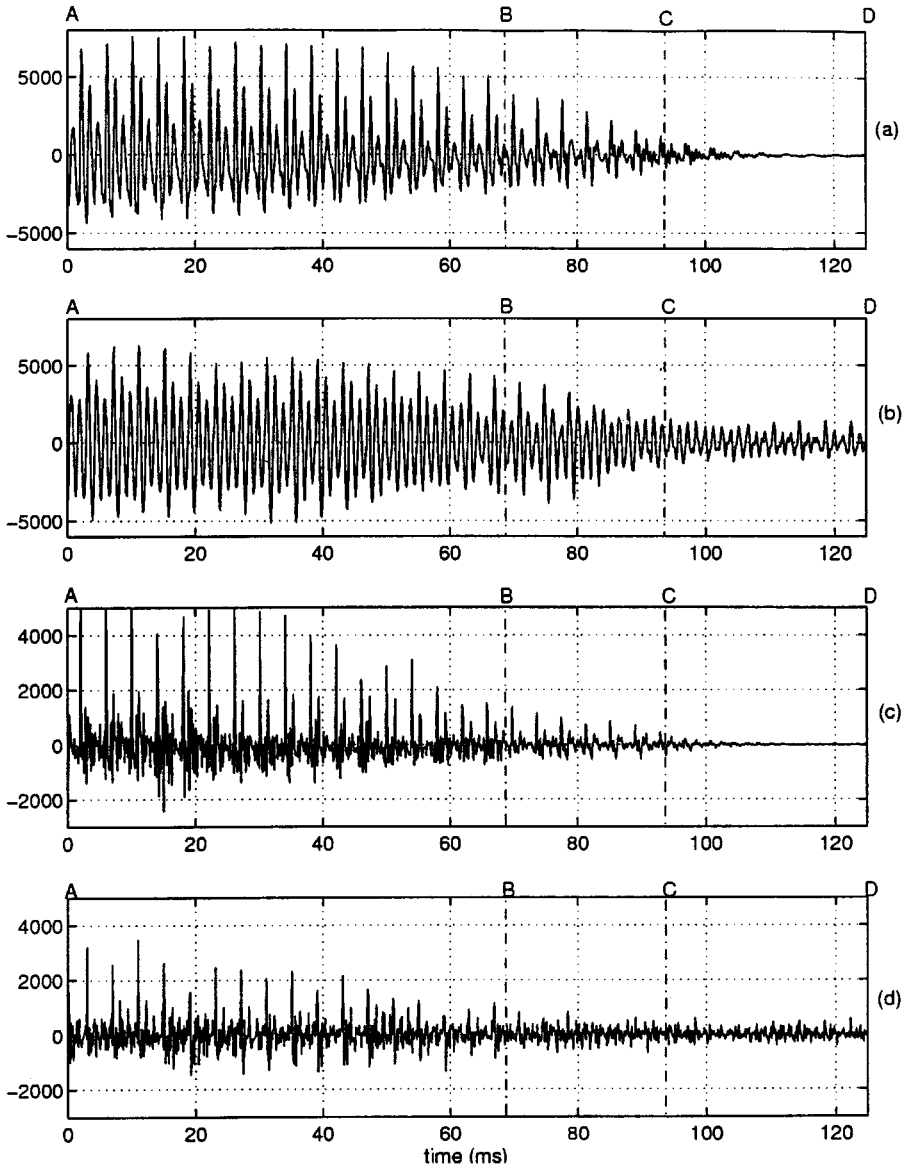


Fig. 2. Comparison of clean and reverberant speech signals: (a) clean speech, (b) signal corrupted by reverberation, (c) LP residual signal for the clean speech in (a), and (d) LP residual signal for the reverberant speech in (b).

$$\text{Noisy speech:} \quad y(n) = s(n) + z(n) \tag{2}$$

where

- $s(n)$ clean speech signal;
- b_k relative amplitude of the reflection arriving after a delay of n_k samples;
- N number of such reflections;
- $z(n)$ additive noise component.

In each model, the first term on the right hand side is the signal component and the second term is the component due to degradation. The main difference between these two models is that, in the case of reverberation, the degrading component is dependent on previous speech data, whereas in the case of noisy speech the degrading component is independent of speech. That is, in the reverberation the degrading component is speech-like.

The relative strength of the reverberant component over the direct component depends on the energy of the speech signal in a short segment around the current instant. This strength can be called signal-to-reverberant component ratio (SRR) at that instant. Likewise, the ratio of the signal energy to the noise energy in a short segment around the current instant is called SNR at that instant. To study the characteristics of SRR and SNR as a function of time, these ratios are computed for short (2 ms) segments of degraded speech. Due to nonstationary nature of speech, the signal energy varies with time. Fig. 1(a) shows a clean speech signal. The energy of the clean speech and the SRR for the reverberant speech are computed for every 2 ms frame shifted by one sample (8 kHz sampling rate) and are shown in Fig. 1(b) and (c), respectively. The reverberant speech signal is generated by convolving the clean speech signal in Fig. 1(a) with the impulse response of a room collected in a normal room at a distance of 1.5 m from the source. Likewise, the SNR is computed for speech degraded by additive noise (overall SNR = 10 dB) and is plotted in Fig. 1(d). In both cases, it is obvious that SRR and SNR vary with time, since the signal energy is also a function of time. In fact, in the case of reverberant speech, both

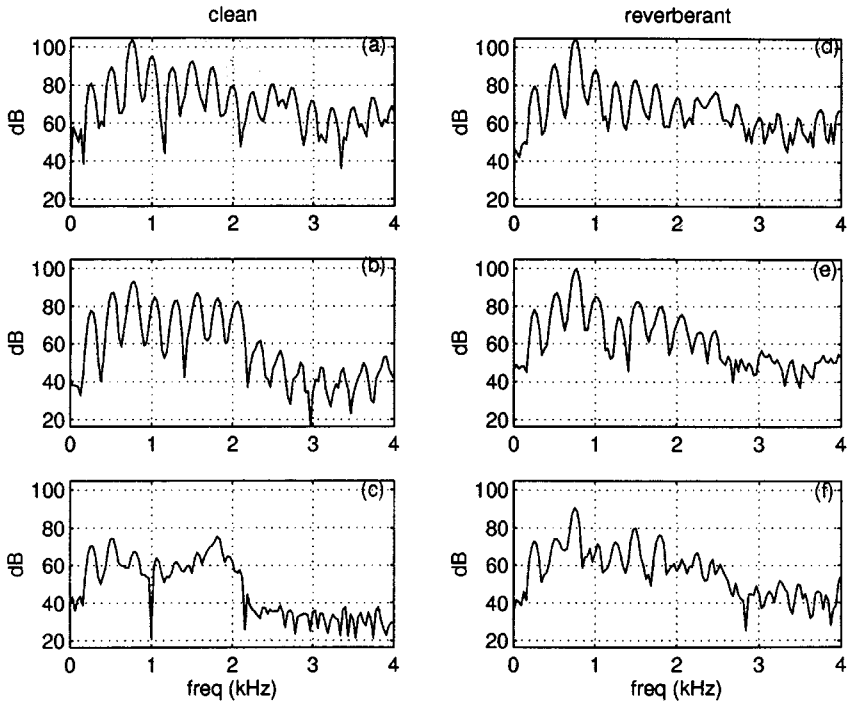


Fig. 3. Comparison of short-time spectra for clean and reverberant speech in different segments. (a)–(c) Short-time spectra of the clean signal in Fig. 2(a) in the regions AB, BC, and CD, respectively. (d)–(f) Short-time spectra of the reverberant signal in Fig. 2(b) in the regions AB, BC, and CD, respectively.

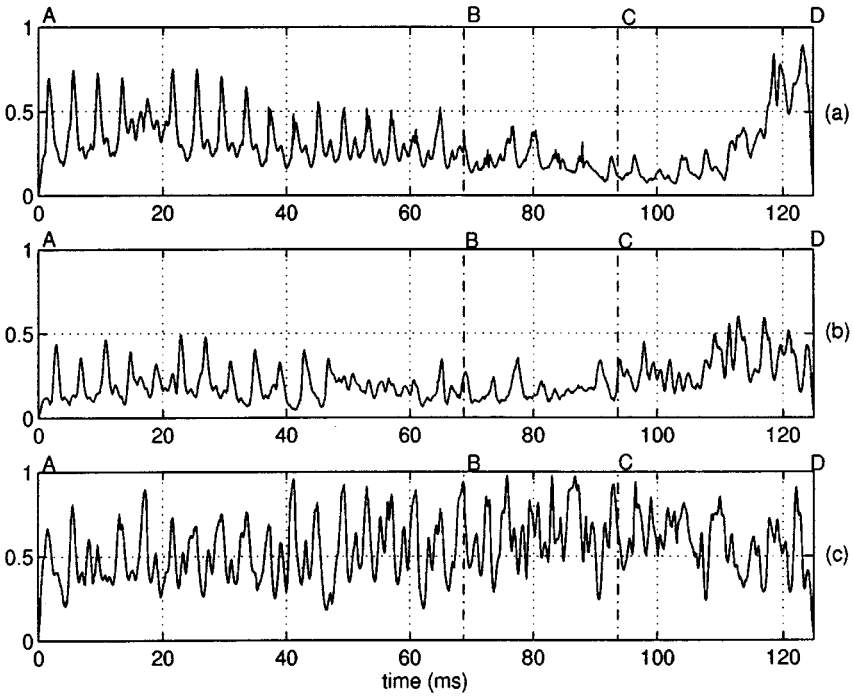


Fig. 4. Comparison of normalized prediction error for: (a) clean, (b) reverberant, and (c) noisy speech (average SNR = 10 dB).

the signal energy and the energy of the degrading component are time-varying, which is not always true in the case of noise-corrupted speech. In Fig. 1(c), we observe that in the 300–400 ms region the SRR is very poor. This is because the direct component is small in this region, whereas there is a large reverberant tail component due to the preceding vowel. In Fig. 1(c) and (d), we also observe that there are finer variations (ripple) in the SRR and SNR plots. This is because of the variation of the signal energy and energy of the degrading component even within a glottal cycle.

The effects of reverberation can be seen by comparing the signal waveforms for clean and reverberant speech signals shown in Fig. 2. The clean speech has damped sinusoidal pattern within each glottal cycle, whereas the reverberant speech is smeared within each cycle [region AB in Fig. 2(b)]. Smearing of the signal within each glottal cycle is more prominent when

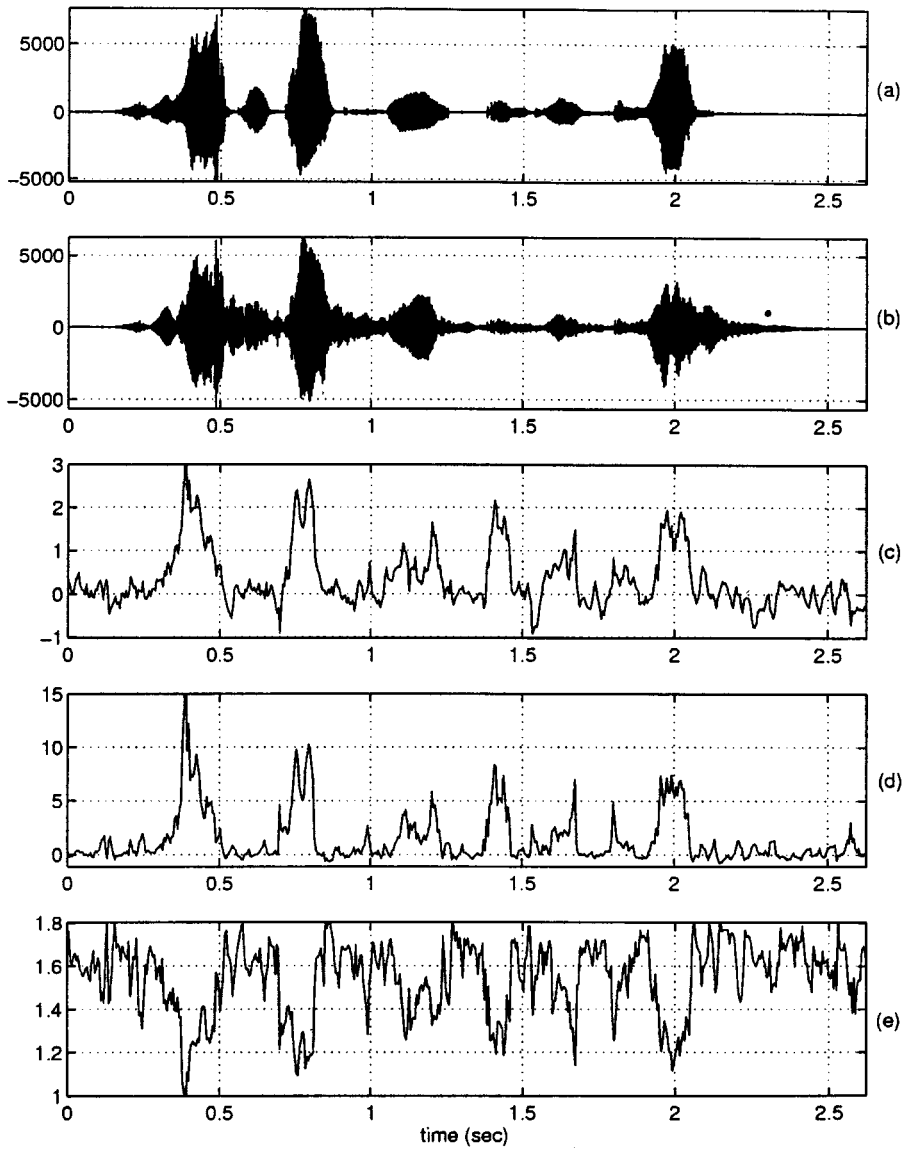


Fig. 5. Characteristics of LP residual signal for reverberant speech: (a) clean speech signal, (b) reverberant speech signal, (c) skewness, (d) kurtosis, and (e) entropy function.

the envelope of the signal waveform is decaying as in the region BC in the figure. The smearing extends for several glottal cycles due to the influence of large amplitude signal component in the region AB. Only the reverberation tail component is present in the low amplitude silence regions (CD).

Nature of the reverberant speech in the spectral domain can be observed by comparing short-time (20 ms) spectra (Fig. 3) for segments in each of the three regions. In all the three cases the dynamic range of the dominant initial portion of the spectral envelope is higher for the reverberant speech compared to that of the clean speech. Thus, there is reduction in the flatness of the spectral envelope due to reverberation. The figure also illustrates that the spectral features of the clean speech are altered significantly due to reverberation, especially for the segments in the regions BC and CD in Fig. 2.

Effect of reverberation can also be seen clearly in the LP residual signal waveform. Fig. 2(c) and (d) shows the LP residual signals for clean and reverberant speech. The residual signal is computed for a segment of 2 ms at every sampling instant, using a fifth-order autocorrelation LP analysis. The residual signal for reverberant speech signal clearly shows that there is a significant direct component of the signal in the reverberant speech in the region AB. This is because for the segments in the region AB the signal amplitudes at the epochs (instants of glottal closure) are higher than the signal amplitudes in the rest of the glottal cycle, like in the case of clean speech. This shows that there are segments in the reverberant speech where the direct component is significantly higher than the reverberant component. In the region BC, due to the decay of the overall signal amplitudes, the reverberation effects of the preceding speech dominate over the direct component. In the region CD the residual signal is mainly due to reverberation.

Comparing the residual signals for clean and reverberant speech signals, the effects of reverberation can be seen within each glottal cycle since the residual signal is much higher in between two epochs when the reverberant component domi-

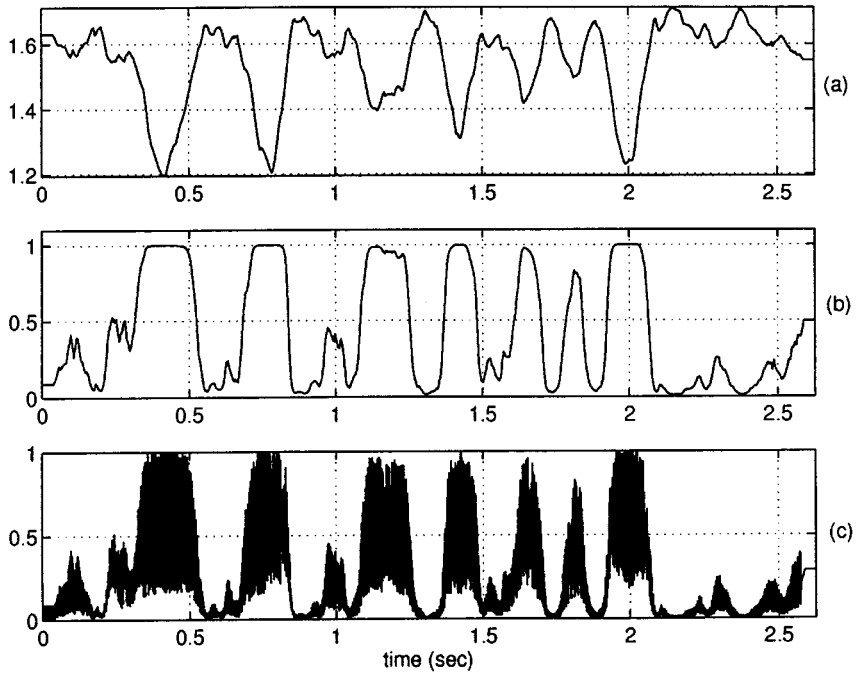


Fig. 6. Various stages in the derivation of the weight function for the LP residual signal: (a) smoothed entropy function, (b) gross weight function, and (c) overall weight function.

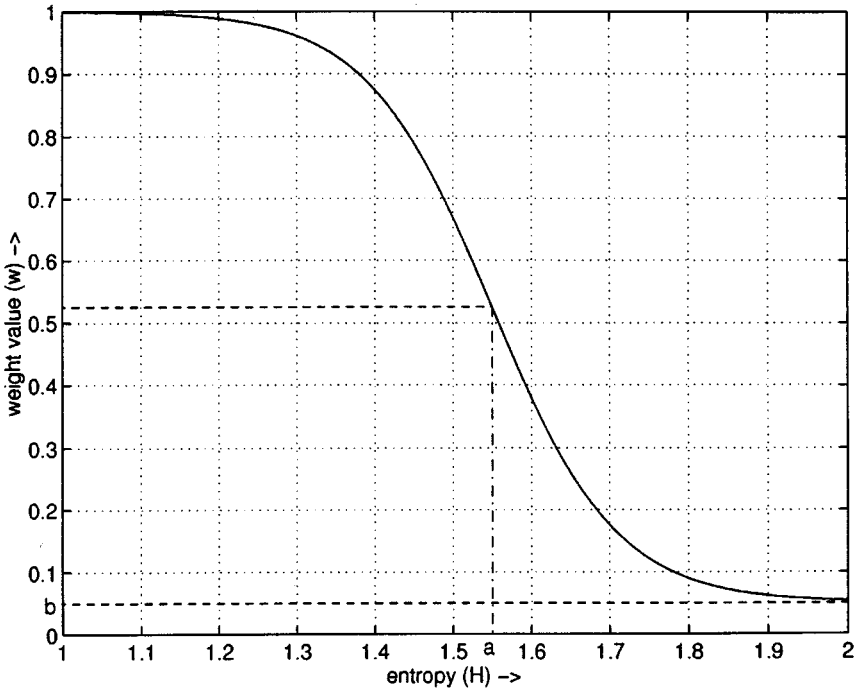


Fig. 7. Mapping function to generate the weight values from the entropy values. The mapping functions $w = ((1-b)/2)\ \tanh(-\alpha_g \pi(H-a)) + ((1+b)/2)$ is shown for $\alpha_g = 1.5$, $a = 1.55$ and $b = 0.05$.

nates. Whenever the direct component of speech is higher than the reverberant component, the LP residual signal at the epochs has significant energy around the instants of glottal closure. Fig. 2(c) and (d) shows that there are regions where the direct component is dominant. We need to identify such regions so that the signals in those regions can be processed to enhance the direct component over the reverberant component. Note that there is no clear evidence of the direct component in the region BC, and there is only reverberant component in the region CD. So the signals in the regions BC and CD need to be attenuated relative to the signal in the region AB. Within the region AB, the signal around the instants of glottal closure need to be enhanced compared to the signal in the rest of the glottal cycle.

First of all, it is necessary to identify these three different regions in the reverberant speech. For this purpose let us observe some more characteristics of the reverberant speech. Fig. 4

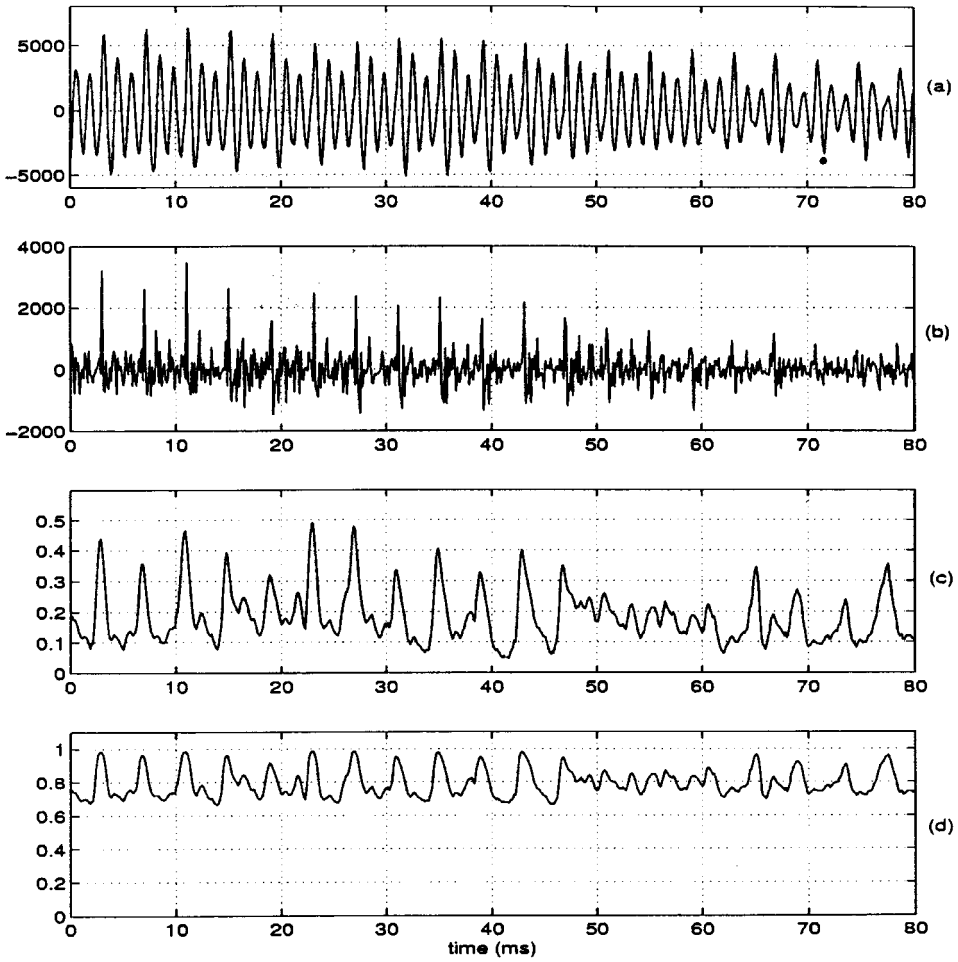


Fig. 8. Derivation of the fine weight function: (a) segment of reverberant speech, (b) LP residual signal, (c) normalized prediction error, and (d) fine weight function.

shows the normalized error (η) of clean and reverberant speech, computed at every sampling instant using a fifth-order autocorrelation LP analysis using a frame size of 2 ms. The normalized errors for both the clean and reverberant speech are similar in the high SRR regions. But the normalized error for the reverberant speech is generally lower than for the clean speech. This is due to the multiplicative effect of the frequency response of the room on the speech spectrum. Multiplication of two spectra produces larger dynamic range and hence reduces the spectral flatness.

In contrast, the speech corrupted by additive noise has higher spectral flatness compared to the clean speech. Thus, the normalized error for the additive noise case is higher than for the clean speech as shown in Fig. 4. Although the LP residual signal for noisy and reverberant speech look similar, their spectral flatness characteristics are distinct. Reverberation decreases the spectral flatness of speech whereas additive noise increases the spectral flatness. In fact, the increase in spectral flatness for additive noise was exploited for developing a method for enhancement of noisy speech [21].

A closer examination of the normalized error plot within each glottal cycle shows that the error is maximum just before glottal closure. This is because the speech signal amplitude is low in this region. The points of maximum η within each glottal cycle can be identified in the high SRR regions such as AB in Fig. 4. It is difficult to see the distinction between open and closed glottis regions in the low SRR regions such as BC. The normalized error in the purely reverberant region (CD) does not show any periodic peaks.

The above study of the characteristics of reverberant speech suggest that we need to address the following issues for enhancement.

1) Which domain to process; temporal or spectral? Which signal to manipulate; original or residual?
2) How to identify the high SRR regions in short (2 ms) segments as well as in the long segments such as AB, BC, and CD?
3) How to process the signal in each of these regions so that the SRR is increased at the fine level (2 ms) within a glottal cycle, and at the gross level (>20 ms segments) as in the regions AB, BC, and CD?
4) How to increase the spectral flatness to the levels of clean speech signal by increasing the normalized error in each segment of speech?
5) How to measure the enhancement realized by a processing method?

In Section III, we discuss some approaches to deal with each one of these issues, and present a method for processing rever-

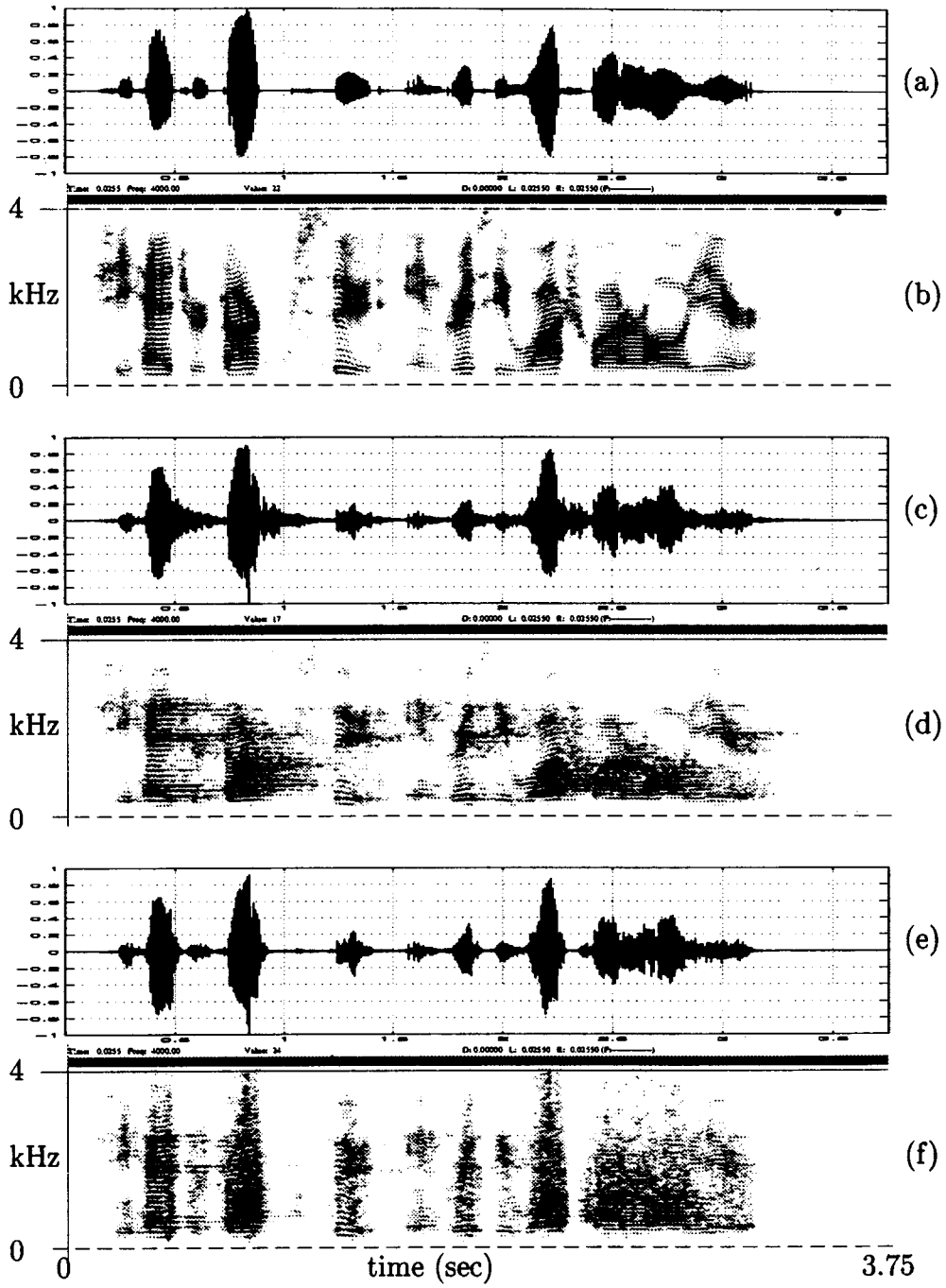


Fig. 9. Results of enhancement of reverberant speech of a male voice: (a) clean speech, (b) spectrogram of clean speech, (c) speech degraded by reverberation, (d) spectrogram of speech degraded by reverberation, (e) speech processed using the proposed algorithm, and (f) spectrogram of processed speech.

berant speech for enhancement. The important point to be noted is that for enhancement of degraded speech, different segments need to be processed differently according to the characteristics of speech in the temporal and short-time spectral domains.

III. Processing Reverberant Speech Using LP Residual Signal for Enhancement

For processing reverberant speech for enhancement, we propose manipulation of the LP residual signal in short (2 ms) and in longer (20 ms) segments in a selected manner. The manipulation basically involves weighting the residual signal samples appropriately. Manipulation of the residual signal is more appropriate than the manipulation of speech signal, especially for short (2 ms) 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. It is interesting to note that any distortion caused by processing the residual signal is smoothed out by the all-pole filter used for synthesis.

LP residual signal is computed by performing the LP analysis on short (2 ms) segments of speech data around every sampling instant. Differenced speech signal samples are used to perform

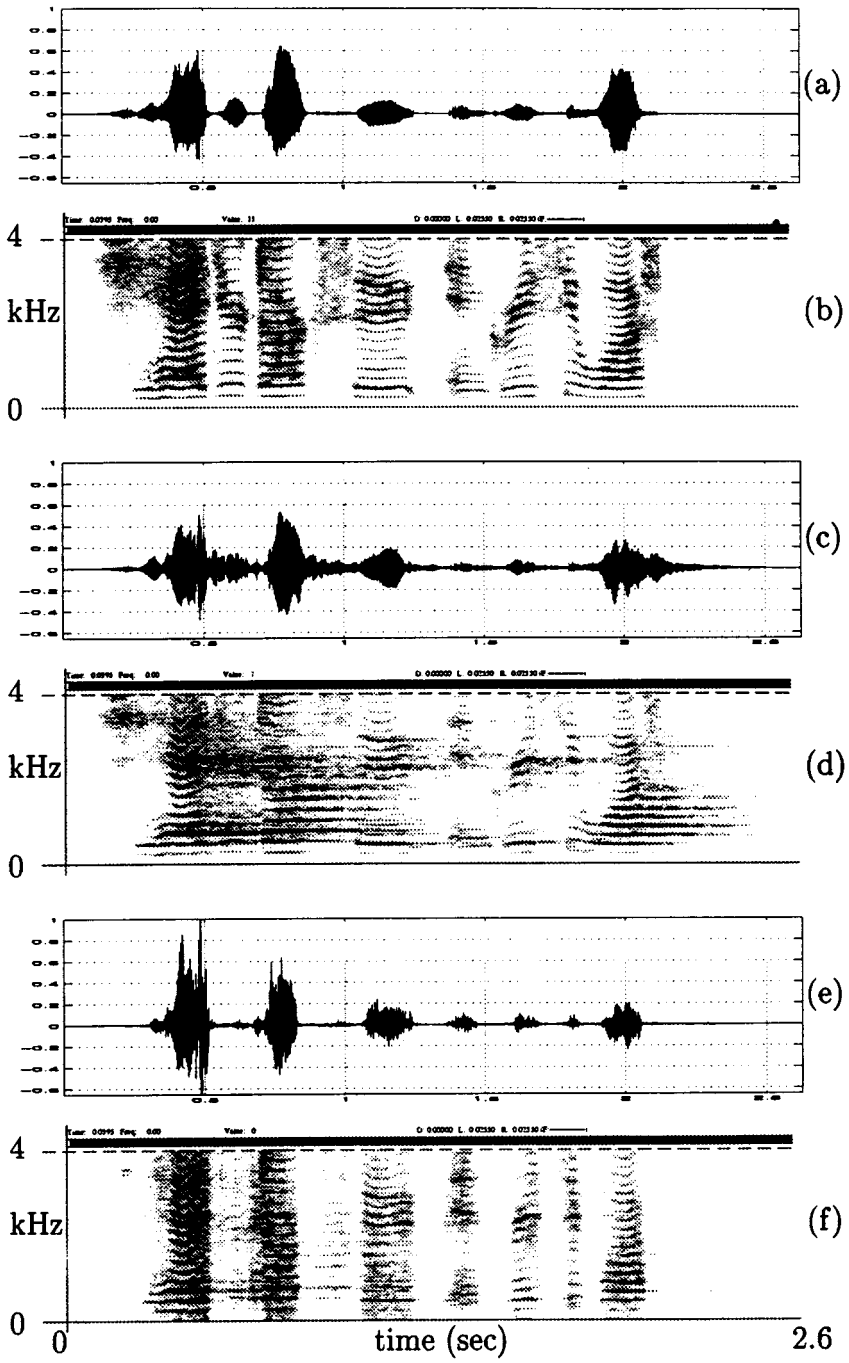


Fig. 10. Results of enhancement of reverberant speech of a female voice: (a) clean speech, (b) spectrogram of clean speech, (c) speech degraded by reverberation, (d) spectrogram of speech degraded by reverberation, (e) speech processed using the proposed algorithm, and (f) spectrogram of processed speech.

TABLE I
ATTRIBUTES OF THE FIVE-POINT SCALE USED FOR SUBJECTIVE EVALUATION

Points	Perceived quality	Level of degradation
5	excellent	imperceptible
4	good	just perceptible but not annoying
3	fair	perceptible, slightly annoying
2	poor	annoying, not objectionable
1	unsatisfactory	very annoying and objectionable

the LP analysis. The LP residual signal is obtained by inverse filtering the speech signal using the LPC's. The reduction of correlation achieved by the inverse filtering is useful to modify the residual signal.

As mentioned earlier, processing of the LP residual signal involves determination of suitable weight function for the residual signal. The weight function is derived for modifying the residual signal both at the fine (within glottal cycle) level and at the gross level. To derive the weight function we need to identify the different SRR regions at the fine and gross levels from the reverberant speech signal. That is, we need to determine the three types of regions such as AB, BC, and CD shown in Fig. 2, and also the regions around the instants of glottal closure in AB. These regions can be identified using the properties of the LP

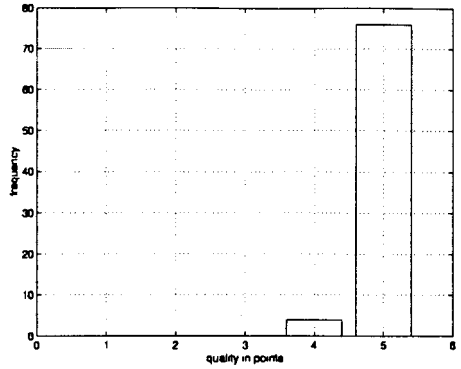


Fig. 11. Frequency histogram showing the frequency distribution of the scores given to the quality of the clean speech signals on a 1–5 point scale.

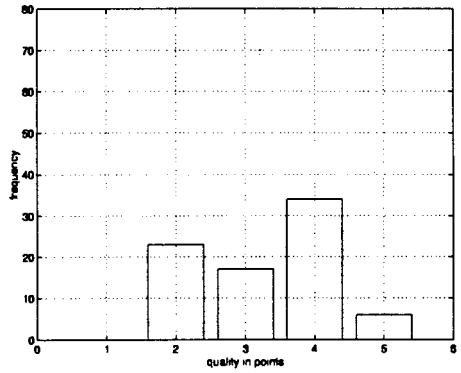


Fig. 12. Frequency histogram showing the frequency distribution of the scores given to the quality of the reverberant speech signals on a 1–5 point scale.

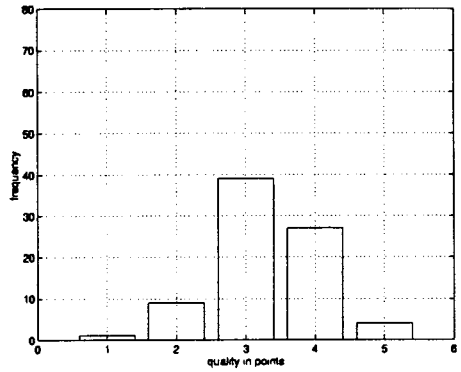


Fig. 13. Frequency histogram showing the frequency distribution of the scores given to the quality of the processed speech signals on a 1–5 point scale.

residual signal for reverberant speech. The regions at the gross level are determined using the statistics of the LP residual signal. In the high SRR regions, the entropy of the distribution of the samples in the LP residual signal is low compared to the entropy in the low SRR regions. This is because the LP residual signal samples exhibit a Gaussian-like probability density function in the reverberant tail regions, and hence the entropy is high. In the high SRR regions, especially in the voiced regions, the peaks in the LP residual signal due to strong excitations of the vocal tract system produce a skewed density function, and hence the resulting entropy is low. To compute the entropy, the probability density function of the samples in each 20 ms segments of the LP residual signal is estimated. A longer (20 ms) segment is

TABLE II
WEIGHTED ITAKURA DISTANCES COMPUTED BETWEEN THE CLEAN SPEECH AND THE REVERBERANT SPEECH ($\bar{d}_{AB}$), AND THE CLEAN SPEECH AND THE PROCESSED SPEECH ($\bar{d}_{AC}$) FOR TEN SPEAKERS

Speaker/Gender	$\bar{d}_{AB}$	$\bar{d}_{AC}$
female#1	942.98	793.82
female#2	1828.30	1315.24
female#3	1145.32	859.12
female#4	2010.83	1663.22
male#1	1434.57	1233.14
male#2	1097.41	1090.46
male#3	889.39	830.90
male#4	2713.85	2330.37
male#5	875.39	663.10
male#6	2811.29	2290.50

used to obtain a good estimate of the histograms of the samples and hence their probability density function. The entropy H_k for the kth frame is given by the following expression [22]:

$$H_k = -\sum_{i=1}^{M} p_i \log(p_i) \tag{3}$$

where p_i is the estimated probability for the ith bin of the histogram, and M is the number of bins in the histogram. The number of bins (M) can be chosen to be in the range 5–20, making sure that there are enough LP residual signal samples per bin. We have chosen a value of $M = 7$. This ensures that there are on an average about 20 samples per bin in each 20 ms frame. The entropy is computed for a 20 ms frame at every 10 ms. Fig. 5(a) and (b) show the clean and reverberant speech signals, respectively. Fig. 5(c) and (d) show the skewness and kurtosis computed for a 20 ms frame of the LP residual signal at every 10 ms. Fig. 5(e) shows the entropy function. It is clear from the figure that both the skewness and kurtosis are high in the regions where the direct component of the signal is strong and so the corresponding entropy is low. The skewness and kurtosis assume values close to zero in the silence and reverberation tail regions because the shape of the estimated probability density function is Gaussian-like [23]. Therefore the entropy in these regions is high as shown in Fig. 5(e).

The entropy function is smoothed by repeating each entropy value in Fig. 5(e) 80 times (corresponding to 10 ms at 8 kHz sampling rate), and smoothing the resulting function using a 600-point mean smoothing filter. From the smoothed entropy function [Fig. 6(a)] a gross weight function w_n^{gross} [Fig. 6(b)] is derived using the nonlinear mapping function shown in Fig. 7. The objective of the nonlinear mapping function is to enhance the contrast between the strong direct speech component and the reverberant component. The values of a and b in Fig. 7 can be varied to derive a suitable mapping function, although the setting of these thresholds is not critical. The entropy function is preferable to the skewness and kurtosis functions for deriving the gross weight function. This is because the entropy function detects even weak speech regions (both voiced and unvoiced) while the skewness and kurtosis functions were found to be sensitive to only the strongly voiced regions.

From the gross weight function [Fig. 6(b)], the three different types of SRR regions can be identified. The regions of rising and

TABLE III
ALGORITHM FOR PROCESSING REVERBERANT SPEECH FOR ENHANCEMENT

Computation of the gross weight function

- Calculate the linear prediction (LP) residual signal using a speech frame of size 20 ms, Hamming window and a 10th order LP analysis by autocorrelation method.
- Block the LP residual signal into 20 ms frames with 10 ms overlap. Compute an M–bin ($M = 7$) histogram of the samples in each frame of the LP residual signal.
- Compute the entropy $H_k = -\sum_{i=1}^{M} p_i \log(p_i)$ for the kth frame, where p_i is the estimated probability in the ith bin of the histogram.
- Compute a smoothed entropy function H_n^s by repeating each entropy value H_k 80 times (corresponds to a frame shift of 10 ms at 8 kHz sampling) and smoothing it with a 600–point mean smoothing filter. This generates a smoothed entropy value at every sampling instant.
- Compute the gross weight function by mapping the smoothed entropy values to weight values using the function

$$w_n^{gross} = \left(\frac{w_{max}^{gross} - w_{min}^{gross}}{2}\right) \tanh\left(-\alpha_g \, \pi \, (H_n^s - a)\right) + \left(\frac{w_{max}^{gross} + w_{min}^{gross}}{2}\right)$$

where w_n^{gross} is the weight value for sampling instant n, w_{max}^{gross} (= 1) is the maximum weight value, w_{min}^{gross} (= 0.05) is the minimum weight value (denoted as b in Fig. 7), α_g (=1.5) is a positive constant which decides the slope of the weight function, a (=1.55) is the entropy value about which the *tanh* function is anti–symmetric and H_n^s is the smoothed entropy at the sampling instant n.

Computation of the fine weight function

- Calculate the normalized LP error for every sample of the differenced speech signal using a frame of duration 2 ms and 5th order LP analysis using the autocorrelation method.
- Remove the trend in the normalized LP error by smoothing it with a 10 ms Hamming window and subtracting the smoothed function from the normalized LP error. The resulting detrended error function η_n is mapped using the nonlinear function

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

where w_n^{fine} is the weight value for sampling instant n, w_{max}^{fine} (= 1) is the maximum weight value, w_{min}^{fine} (= 0.6) is the minimum weight value, α_f (=1.5) is a positive constant which decides the slope of the weight function and η_n is the detrended error value at the sampling instant n.

Synthesis of enhanced speech

- Compute the overall weight function by multiplying the gross and fine weight functions.
- LP residual signal is derived for every sample using 2 ms frames. The residual signal is multiplied with the overall weight function. The weighted residual signal is passed through the time–varying LP all–pole filter to obtain enhanced speech. At each sampling instant the LPCs are given by $a_{kn} r_n^{-k}$, where a_{kn} is the kth LPC at instant n. The damping factor r_n, restricted to the range 0.9–1.0, is derived using a linear map of the fine weight function.

high values of the weight function correspond to the high SRR regions (like region AB in Fig. 2). The falling portions correspond to the low SRR regions (like region BC in Fig. 2). The low weight function regions correspond to the reverberant component regions (like region CD in Fig. 2). To derive the fine weight function, the normalized error (η) is computed at each sampling instant using a frame size of 2 ms and a fifth-order LP analysis. The normalized error is shown in Fig. 8(c) for a segment of 80 ms of speech shown in Fig. 8(a). The peaks in the error function generally correspond to the region around the glottal excitation points, at which the LP residual signal [Fig. 8(b)] also has large amplitudes. Note that the normalized LP error shows the characteristic peaks in the initial 50 ms segment because of the strong direct component. These peaks are not prominent in the latter 30 ms segment because of the stronger reverberant component. A second weight function, which we refer to as the fine weight function, is derived from the normalized error by removing the global trend in the normalized error function and then mapping it using the following function:

$$w_n^{\text{fine}} = \left(\frac{w_{\max}^{\text{fine}} - w_{\min}^{\text{fine}}}{2}\right) \tanh\left(\alpha_f \pi \eta_n\right) + \left(\frac{w_{\max}^{\text{fine}} + w_{\min}^{\text{fine}}}{2}\right) \tag{4}$$

where

w_n^{fine}	weight value at the sampling instant n;
$w_{\max}^{\text{fine}} (=1)$	maximum weight value;

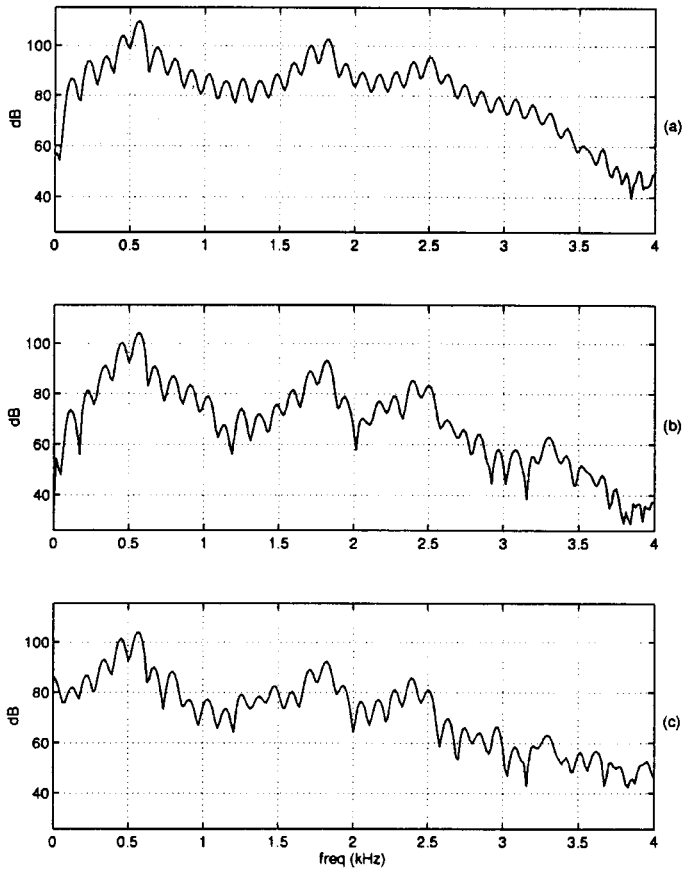


Fig. 14. Short-time spectra of a segment of speech for: (a) clean speech signal, (b) reverberant speech signal, and (c) processed speech signal.

$w_{\min}^{\text{fine}}$	minimum weight value;
$\alpha_f\ (=1.5)$	positive constant which decides the slope of the weight function;
η_n	detrended normalized error value at the sampling instant n.

The fine weight function for the segment of the signal in Fig. 8(a) is shown in Fig. 8(d). The fine weight function provides relative weighting of short segments within a glottal cycle in the high SRR regions. The overall weight function [Fig. 6(c)] is obtained by multiplying the gross weight function with the fine weight function. The overall weight function and the LP residual signal are multiplied to derive a modified residual signal. The modified residual signal is used to excite the fifth-order all-pole filter to obtain enhanced speech. The filter is updated at every sampling instant.

A comparison of the clean speech waveform and reverberant speech waveform in the voiced regions shows that within a glottal cycle the reverberant speech waveform does not decay as rapidly as the clean speech waveform. Despite the deemphasis of low SRR regions within a glottal cycle by the fine level weight function, the decay of the envelope within a glottal cycle is not restored in the processed speech waveform. Hence, there is a need to increase the flatness by manipulating the spectrum. One way of doing this is to modify the filter coefficients to $a_{kn}r_n^{-k}$ for $k = 1, 2, \cdots, p$, where p is the order of the all-pole filter, $r_n < 1$ and a_{kn} is the kth LPC at the sampling instant n. The damping factor r_n at each sampling instant is varied according to the value of the fine weight function. The value of r_n is restricted to the range 0.9–1.0. The modification of LPC's will enable the roots of the all-pole filter move closer to the origin in the z-plane. Due to dependence of r_n on the fine weight function, the proposed modification of LPC's is equivalent to damping the resonances of the vocal tract system toward the end of the glottal cycle. The algorithm for processing reverberant speech for enhancement is given in Table III.

IV. Experimental Results

In this section the performance of the proposed method is examined for processing speech data collected under reverberant conditions. The performance of the method is illustrated through spectrographic results, subjective and objective evaluations. For this purpose the speech data was collected in an empty room of dimensions $2.5\ \text{m} \times 1.8\ \text{m} \times 2.7\ \text{m}$. The microphone was placed about 1.5 m away from the speaker. In all the experimental results presented in this section, the preemphasized speech signal was processed using the algorithm given in Table III.

The results of enhancement of a speech signal corresponding to the sentence "She had your dark suit in greasy wash water all year" uttered by a male speaker are given in the waveform and spectrographic plots in Fig. 9. The utterance is taken from the TIMIT database [24]. Fig. 9(a) and (b) show the clean speech signal and its spectrogram, respectively. Fig. 9(c) and (e) show the reverberant and processed speech signals and Fig. 9(d) and (f), the corresponding spectrograms, respectively. From the spectrograms it is evident that the effects of reverberation (e.g., the reverberation tails) are significantly reduced. The performance of the method was tested for female voice also. The resulting signal waveforms and spectrograms are shown in Fig. 10. Here also the signal corresponds to the sentence "She had your dark suit in greasy wash" taken from the TIMIT database.

Subjective tests were conducted to study the improvement in quality of the processed speech. Perceptually, the processed signal sounds less reverberant than the unprocessed one. The subjective evaluation was done by eight listeners who are students in the age group of 21–25 years. The listeners are fluent in English but are not experienced in subjective evaluation. The evaluation was done on a set of ten different sentences uttered by ten different speakers taken from the TIMIT database. The test set comprised of four female voices and six male voices. The clean speech, reverberant speech and processed speech were played to the listeners in that order. They were asked to grade the quality of each of the three speech signals on the following five-point scale (see Table I). The grades given to the clean, reverberant and processed speech signals are shown as frequency histograms in Figs. 11–13, respectively. It is clear from Fig. 12 that the listeners were not consistent in judging the quality of the reverberant speech signals. However, the perceived quality of the processed speech signals was consistent. The mean opinion score (MOS) [25] for the processed speech was found to be 3.30 on the five point scale, while it was 3.28 for the reverberant speech. The MOS does not show any significant improvement due to processing. This is partly due to the listener's

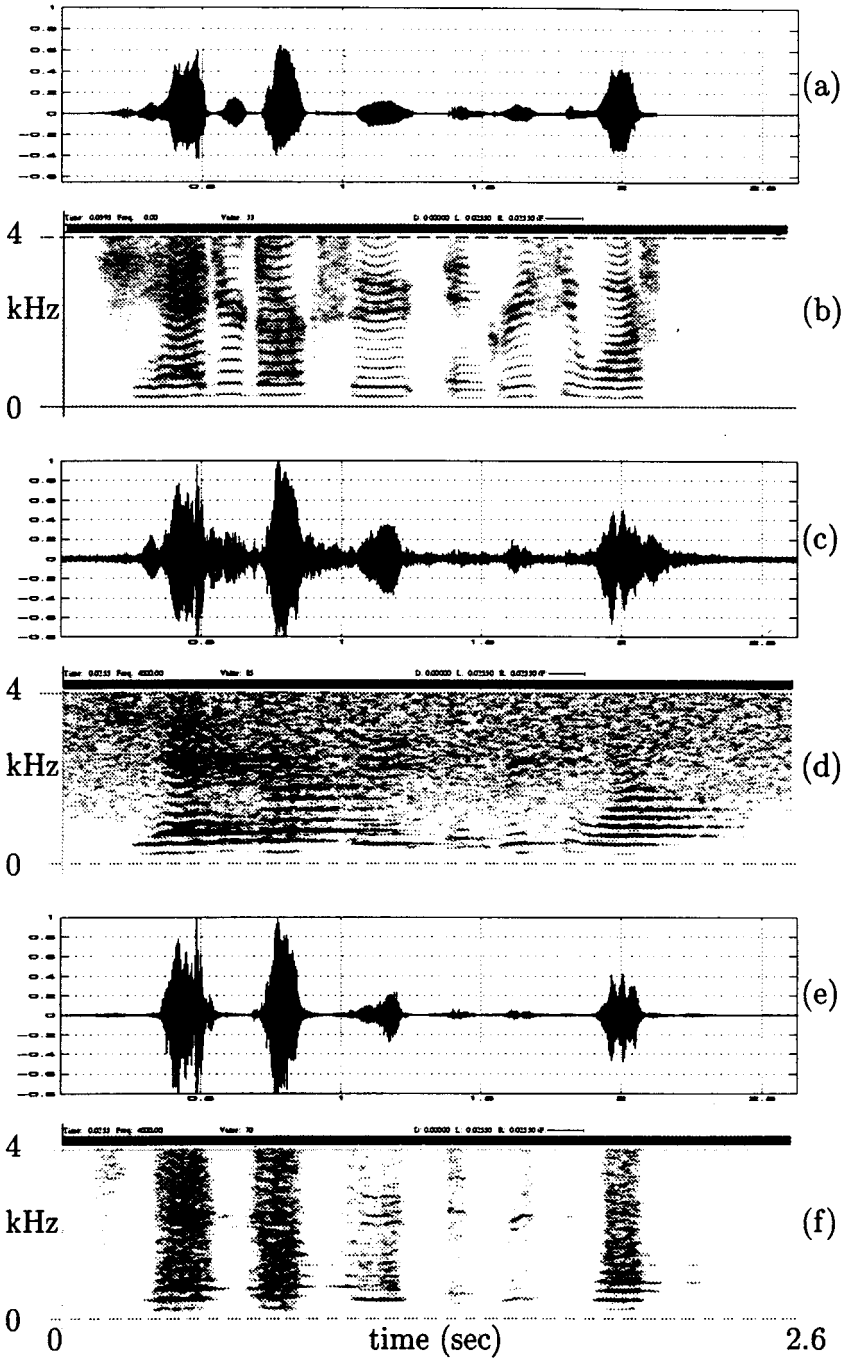


Fig. 15. Results of enhancement of speech degraded by reverberation and noise: (a) clean speech, (b) spectrogram of clean speech, (c) speech degraded by reverberation and noise (SNR = 20 dB), (d) spectrogram of speech degraded by reverberation and noise, (e) speech processed using the proposed algorithm, and (f) spectrogram of processed speech.

inability to make subjective assessment of the quality. The results for different values of the parameters used in the algorithm show that the parameter settings are not very critical. They provide some tradeoff between quality and enhancement in the processed signal.

The improvement due to processing is illustrated more clearly by the weighted average of Itakura distances obtained between the clean speech and the reverberant speech, and the clean speech and the processed speech. It is important to note that Itakura distance emphasizes the differences in the spectral peaks of two spectra than the differences in the spectral valleys. Thus it has a good correlation to the subjective quality [25].

The data used is the same ten speaker corpus used for the subjective evaluation. The clean speech, reverberant speech and the processed speech for each speaker were first time aligned. The Itakura distance [26] $d_{k\mathrm{AB}}$ was computed between each pair of corresponding frames k in the clean speech and reverberant speech signals. The weighted average $\overline{d}_{\mathrm{AB}}$ of these distances was obtained using

$$\overline{d}_{\mathrm{AB}} = \sum_{k} w_k^{\mathrm{gross}} d_{k\mathrm{AB}} \tag{5}$$

where the frame weights w_k^{gross} were derived from the gross weight function w_n^{gross}. The frame weighting is done to deemphasize the contribution of the Itakura distances in the pause regions of the speech signal. Similarly, the weighted average $\overline{d}_{\text{AC}}$ of the distances between the clean speech and processed speech signals was obtained. The distances $\overline{d}_{\text{AB}}$ and $\overline{d}_{\text{AC}}$ for each of the ten speakers are given in Table II. The distance $\overline{d}_{\text{AB}}$ is consistently greater than the distance $\overline{d}_{\text{AC}}$. This is because the dynamic range of the linear prediction spectra of the processed speech signal is lower compared to that of the reverberant speech. Therefore, the LP spectra of processed speech yield lower distances with those of clean speech, compared to the distances of reverberant speech. This is also illustrated in Fig. 14. The figure shows the short-time (20 ms) spectra for a voiced segment of speech for clean, reverberant and processed speech. The reduction in the dynamic range of the spectra after processing can be seen clearly, especially around the formant regions. Thus, the spectral flatness of the clean speech is restored to some extent. For enhancement of noisy speech, on the other hand, one attempts to lower the spectral flatness by increasing the spectral dynamic range [21].

In a practical speakerphone-like situation, in addition to degradation due to reverberation, there will be ambient noise also. Fig. 15 shows this situation. The reverberant speech signal in Fig. 10(c) is corrupted by additive random noise so that the overall SNR is 20 dB. The noise added reverberant speech is shown in Fig. 15(c). The processed speech signal is shown in Fig. 15(e). The spectrograms for Fig. 15(c) and (e) are shown in Fig. 15(d) and (f), respectively. The improvement can be clearly seen in the spectrogram in Fig. 15(f). We observe that in the silence regions the noise level as well as the reverberation are significantly reduced. This is because the noise increases the randomness in the LP residual signal, more so in the silence regions and hence increases the entropy. Hence, in the silence regions both the reverberation tails and the noise increase the entropy. Thus, the gross weight function will have small values in the silence regions. We also observe from the processed signal in Fig. 15(e) and the spectrogram in Fig. 15(f) that the weak signal segments are severely attenuated, which produces some distortion in the processed speech signal.

There will be some reduction in quality when the proposed algorithm is applied to clean speech. But this reduction in quality is offset by the advantage due to enhancement obtained in processing degraded speech.

V. Conclusions

In this paper, we have presented a new approach for processing reverberant speech. The proposed method is based on the knowledge that the speech signal energy fluctuates over a large dynamic range in short segments (2 ms). Thus, the SRR varies significantly over different segments of speech. By identifying the high SRR regions, and enhancing such regions at gross level and at fine (within glottal cycle) level one can achieve enhancement of reverberant speech. The processing was done by weighting the LP residual signal, and the weight function was derived using the characteristics of the reverberant speech in different regions. The resulting signal shows reduction in the perceived reverberation without significantly affecting the quality. By adjusting the parameters used for obtaining the weight function, the comfort level in the processed signal can be traded with the distortion caused by the manipulation. Thus processing the LP residual signal provides an alternative approach for enhancement of reverberant speech. A uniform approach for processing reverberant speech as in [14]–[16] may not be satisfactory, since the reverberation affects the speech differently in different segments due to nonstationary nature of the speech signal.

The key ideas in this paper are as follows:

1) need to process different regions of reverberant speech differently;
2) advantage of manipulating the residual signal samples for enhancement;
3) ability to tune the processing depending on the level of tolerance of distortion versus the desired comfort level.

It is interesting to note that only regions of high SRR need to be processed for enhancement, whereas the low SRR and the reverberant tail regions should be deemphasized to obtain perceptually significant enhancement.

References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 113–120, Apr. 1979.

[2] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech, Audio Processing*, vol. 3, pp. 251–266, July 1995.

[3] M. Tohyama, R. H. Lyon, and T. Koike, "Pulse waveform recovery in a reverberant condition," *J. Acoust. Soc. Amer.*, vol. 91, pp. 2805–2812, May 1992.

[4] A. P. Petropulu and S. Subramaniam, "Cepstrum based deconvolution for speech dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Adelaide, Australia, Apr. 1994, pp. 9–13.

[5] S. Subramaniam, A. P. Petropulu, and C. Wendt, "Cepstrum-based deconvolution for speech dereverberation," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 392–396, Sept. 1996.

[6] Y. M. Perlmutter, L. D. Braida, R. H. Frazier, and A. V. Oppenheim, "Evaluation of a speech enhancement system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 1977, pp. 212–215.

[7] J. S. Lim, A. V. Oppenheim, and L. D. Braida, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 354–358, Aug. 1978.

[8] M. R. Sambur, "Adaptive noise canceling for speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 419–423, Oct. 1978.

[9] D. Malah and R. V. Cox, "A generalized comb filtering technique for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Paris, France, May 1982, pp. 160–163.

[10] Y. M. Cheng and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence," *IEEE Trans. Signal Processing*, vol. 39, pp. 1943–1954, Sept. 1991.

[11] J. H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 59–71, Jan. 1995.

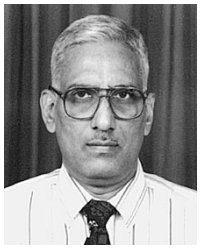
[12] S. Nandkumar and J. H. L. Hansen, "Dual-channel iterative speech enhancement with constraints on an auditory spectrum," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 22–34, Jan. 1995.

[13] M. Tohyama, H. Suzuki, and Y. Ando, *The Nature and Technology of Acoustic Space*. London, U.K.: Academic, 1995.

[14] T. Langhans and H. W. Strube, "Speech enhancement by nonlinear multiband envelope filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Paris, France, April 1982, pp. 156–159.

[15] H. Hirsch, "Automatic speech recognition in rooms," in *Signal Processing—V: Theories and Applications*, J. L. Lacome, A. Chehilian, N. Martin, and J. Malbos, Eds. Amsterdam, The Netherlands: Elsevier, 1988.

[16] C. Avendaño and H. Hermansky, "Study on the dereverberation of speech based on temporal envelope filtering," in *Proc. Int. Conf. Spoken Language Processing*, Philadelphia, PA, Oct. 1996, pp. 889–892.
[17] M. Omologo, P. Svaizer, and M. Matassoni, "Environmental conditions and acoustic transduction in hands-free speech recognition," *Speech Commun.*, vol. 25, pp. 75–95, Aug. 1998.
[18] H. Wang and F. Itakura, "An approach of dereverberation using multimicrophone sub-band envelope estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1991, pp. 953–956.
[19] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 145–152, Feb. 1988.
[20] B. Yegnanarayana, P. Satyanarayana Murthy, C. Avendaño, and H. Hermansky, "Enhancement of reverberant speech using LP residual," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, Seattle, WA, May 1998, pp. 405–408.
[21] B. Yegnanarayana, C. Avendaño, H. Hermansky, and P. Satyanarayana Murthy, "Processing linear prediction residual for speech enhancement," in *Proc. EUROSPEECH'97*, Patras, Greece, Sept. 1997, pp. 1399–1402.
[22] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed, New York: McGraw-Hill, 1991.
[23] W. H. Press, S. A. Teukolsky, W. T. Vellerling, and B. P. Flannery, *Numerical Recipes in C*. New Delhi, India: Cambridge Univ. Press, 1992.
[24] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," in *Proc. DARPA Workshop Speech Recognition*, Feb. 1986, pp. 93–99.
[25] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, New York: Macmillan, 1993.
[26] P. E. Papamichalis, *Practical Approaches to Speech Coding*. Englewood Cliffs, NJ: Prentice-Hall, 1987.



B. Yegnanarayana (M'78–SM'84) was born in India on January 9, 1944. He received the B.E., M.E., and the Ph.D. degrees in electrical communication engineering from the Indian Institute of Science, Bangalore, in 1964, 1966, and 1974, respectively.

He was a Lecturer from 1966 to 1974 and an Assistant Professor from 1974 to 1978 with the Department of Electrical Communication Engineering, Indian Institute of Science. From 1966 to 1971, he was engaged in the development of environmental test facilities for the Acoustic Laboratory, Indian Institute of Science. From 1977 to 1980, he was a Visiting Associate Professor of computer science at Carnegie Mellon University, Pittsburgh, PA. He was a Visiting Scientist at ISRO Satellite Center, Bangalore, from July 1980 to December 1980. Since 1980, he has been a Professor with the Department of Computer Science and Engineering, Indian Institute of Technology, Madras. From July 1994 to January 1995, he was a Visiting Professor with the Institute for Perception Research (IPO), Eindhoven Technical University, Eindohoven, The Netherlands. Since 1972, he has been working on problems in the area of speech signal processing. He is presently engaged in research activities in digital signal processing, speech recognition, and neural networks.

Dr. Yegnanarayana is a Member of the Computer Society of India and a Fellow of the Institution of Electronics and Telecommunications Engineers of India, the Indian National Science Academy, and the Indian National Academy of Engineering.

P. Satyanarayana Murthy was born in Kakinada, India, in 1971. He received the B.E. degree in electronics and communication engineering from Chaitanya Bharathi Institute of Technology, Osmania University, India, in 1992, and the M.Tech. and Ph.D. degrees in electrical engineering from the Indian Institute of Technology, Madras, in 1994 and 1999, respectively.

From January 1994 to July 1994, he was a Senior Project Officer in the Department of Computer Science and Engineering, Indian Institute of Technology, Madras. He is currently a Manager, Research and Development, with Speech and Software Technologies (India) Pvt. Ltd., Madras. His research interest is in speech signal processing.