PROCESSING LINEAR PREDICTION RESIDUAL FOR SPEECH ENHANCEMENT

B. Yegnanarayana †, Carlos Avendano ‡, Hynek Hermansky ‡, and P. Satyanarayana Murthy †

†Department of Computer Science and Engineering
Indian Institute of Technology, Madras 600 036, India
‡Department of Electrical Engineering
Oregon Graduate Institute of Science & Technology
Portland, Oregon, USA

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 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, and (c) At the short–time spectrum level, by enhancing the spectral peaks over spectral valleys. Processing of noisy speech for enhancement involves mostly weighting the LP residual samples. The weighted residual samples are used to excite the time–varying LP filter to produce enhanced speech.

1. INTRODUCTION

Speech signal collected under normal environmental conditions is usually degraded due to noise and distortions. Performance of speech systems depends critically on the effect of these environmental conditions on the parameters and features extracted from the speech signal [1]. The quality of the recorded speech is also affected significantly due to noise and distortions. 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 [2], [3], [4], [5]. Most of these studies focussed on enhancement based on suppression of noise [2], [4]. These methods disturb the spectral balance in speech, resulting in unpleasant distortions in the enhanced speech. Speech enhancement has also been accomplished by smoothing the temporal contours of the parameters or features, like spectral band energies [6]. The smoothing reduces random fluctuations in the parameter contours caused by noise. The parameters of speech are usually related to short–time spectra, and hence smoothing the temporal variations of spectral features may sometimes introduce unnatural spectral changes which are perceived as distortions in the enhanced speech.

In many of the above mentioned attempts, no effort has been made to study the characteristics of the source signal, like the linear prediction (LP) residual, for example, for enhancement. The primary reason for this is that the residual is an uncorrelated error signal, and hence it is noise–like. It is not expected to have any features useful for speech enhancement. However, 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 objective and scope of our study in this paper. We also discuss the characteristics of noisy speech and the basis for the approach. In Section 3, we develop methods for speech enhancement based on the characteristics of the LP residual of a noisy speech signal. We propose enhancement at three levels, each level improving some feature of speech in the noisy signal, without significantly affecting the quality. In Section 4 we discuss the application of the proposed method for different types of additive noise.

2. BASIS FOR THE METHOD

Humans beings perceive speech by capturing some features from high SNR regions in the spectral and temporal domains, and then interpolate at various levels by capturing features in the low SNR regions. Therefore, speech enhancement should primarily aim at highlighting the high SNR regions relative to the low SNR regions. This relative emphasis of features in the high SNR regions over the features in the low SNR regions should be accomplished without causing distortions in speech, which otherwise may cause annoyance of a type different from that due to additive noise. The objective in this study is to accomplish this enhancement by suitably modifying the source and system features of speech production in the signal.

2.1. Background

Before we proceed to discuss our approach, let us briefly review some characteristics of noisy speech. The 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 is also low prior to the release of a stop sound and in the fricative sounds. Even within a pitch period of voiced sounds, due to the damped sinusoidal nature of the impulse response of the vocal tract system, the signal energy is usually higher in the vicinity of the major excitation of the vocal tract system, which is the instant of glottal closure in each pitch period [7]. Even in the spectral domain, due to the large dynamic range of the speech signal, the spectral levels for large amplitude formants will be typically much higher (20-30 dB) than the levels of low amplitude formants. For a given additive white Gaussian noise (AWGN), the SNR varies as a function of frequency in the spectral domain. Thus SNR is different in different segments of speech in both time and frequency domains.

2.2. Noise in the LP Residual

Typically, noise samples are uncorrelated, whereas speech samples are correlated. In the presence of additive noise, spectral flatness of speech increases, becoming more flat in the low SNR portions of the spectrum. Also, the low amplitude regions in the signal contain samples that are less correlated. Thus, 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 capture the high SNR portions to the extent possible, and enhance them 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.

We will exploit the characteristics of the LP residual of noisy speech to accomplish speech enhancement. We attempt to enhance the residual in the regions around the glottal closure in the voiced speech segments and reduce the energy levels of the residual in the unvoiced and silence regions. By exciting the time–varying LP filter (derived from the noisy speech) with the modified residual we can produce a significantly enhanced speech without causing much distortion.

The LP residual can be derived for the noisy speech using a frame of about 25 ms duration and a frame rate of about 100 frames per second. Note that even in the LP residual of noisy speech, the SNR in different regions remains same as in the noisy speech. Thus SNR as a function of time or frequency is exactly same in both the noisy signal and the residual. Inverse filtering merely reduces the correlation between samples existing in the noisy speech signal. Since the residual samples are uncorrelated, 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 short-time spectral analysis. Thus the truncation effects of the analysis window are significantly reduced.

2.3. Spectral Flatness

For each small segment of the residual signal, the energy ratio of the noisy signal and the corresponding portion of the residual 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. 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) segment will be nearly unity. Otherwise, the ratio will be quite large. Note that for noisy segments this ratio of energies will be nearly unity. Thus the ratio of the energies gives an indication of the signal and noise regions of the signal. Using a 12 th order LP analysis, the ratio of energy values for a 10 dB SNR situation (see Fig. 1(a)) computed for each 2 ms frame is shown in Fig. 1(b). Note that even weak signal regions are discernible in the ratio plots. (The noisy signal in Fig. 1(a) is generated by adding white Gaussian noise to a clean speech signal). The ratio can be interpreted as the inverse of spectral flatness of the noisy signal, the maximum flatness being one, corresponding to the energy ratio of 0 dB.

Because of the uncorrelated nature of the residual samples, these samples can be manipulated to some extent without producing significant distortions in the reconstructed speech [8]. It is this manipulative capability of the residual we would like to exploit for enhancement in our study.

3. MANIPULATION OF LP RESIDUAL

The basic principle of our approach for speech enhancement is to identify the low SNR regions in the LP residual, 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, and the modified residual is used to excite the time-varying LP filter derived from the given noisy speech 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 inverse spectral flatness and the relative residual energies between adjacent frames, and (c) At spectral level to enhance the features in the short–time (10–20 ms) spectrum that could not be affected by the fine level operations.

3.1. Gross Temporal Level

At this level the regions corresponding to low and high SNR regions are identified from the characteristics of the LP residual. A weighting function for the residual samples is derived based on smoothed inverse spectral flatness characteristics of the inverse filtering operation on the noisy speech signal. The inverse flatness characteristics are derived by computing the ratio of the energy in the noisy signal to the energy in the residual in each short interval of about 2 ms. The LP residual signal itself is derived from the noisy speech using a 12 th order LP analysis on each 25 ms frame overlapping by 16 ms. Note that the frame size, frame rate and the LP order are not critical for this study. The ratio gives an indication of the inverse spectral flatness as a function of time. The inverse flatness plot is smoothed using a 17 point Hamming window. The size of the smoothing window is also not very critical. The low SNR noisy regions have an inverse flatness close to unity (0 dB), and the high SNR regions have a high inverse flatness value. A weight function is derived from the smoothed inverse flatness characteristics in such a way that the residual signal regions corresponding to flatness near unity (0 dB) are reduced in energy relative to the regions with high inverse flatness. A mapping function of $tanh(x)$ type can be used to map the smoothed inverse spectral flatness values to the weight values for each short frame of 2 ms residual signal data. The parameters of the function are determined by the overall behavior of the smoothed inverse flatness function. 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 weighting function for each sample of the residual as shown in Fig. 1(c). The weighting function clearly indicates the low and high SNR regions.

3.2. Finer Temporal Level

For voiced segments, if the SNR is low in some short (1-2 ms) segments, then the residual 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 pitch period due to damping of the formants. The fluctuations in the residual energy contour for short (2 ms) segments illustrate the energy differences between adjacent segments. A weighting function at the fine level can be derived from the residual energy plot, by deemphasizing the frames corresponding to the valley frames relative to the peak frames. But for noisy speech, the residual is noisy and so the short–time (2-3 ms) energy of the residual may not be reliable to use as a weight. Hence, the Frobenius norm [9] of the toeplitz matrix (see (1) below) constructed using the noisy speech samples in a frame of 2 ms duration can be used to represent the short–time energy of the corresponding frame of LP residual [10]. Though indirect, this method has the advantage of exploiting the envelope information in the noisy speech waveform. The toeplitz prediction matrix $\mathbf{X}$ is given by

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

where $x_1, x_2, ..., x_M$ are the noisy speech samples in a frame of length M (which is 22 for 2 ms duration at 11 kHz sampling) and p is the linear prediction order (12-14). The Frobenius norm is computed for every sample shift of the 2 ms frame. The weighting function is derived using the adjacent frame differenced log residual energy plot. The weight values are dictated by the amount of change from peak to valley point. The maximum change is restricted to the interval 0.1 to 1.0, although this setting is again not very critical. The finer weighting function derived using this range is shown in Fig. 1(d). The overall weighting function is obtained by multiplying the gross weighting function derived from the smoothed flatness plot with the fine weighting function derived from the residual energy plot. The final weighting function for the residual samples is shown in Fig. 1(e). Enhanced speech is generated by exciting the all–pole filter with this modified residual.

3.3. Spectral Level

The LP filters for segments of size 30 ms, has both high and low SNR regions of the signal. In the reconstruction, even though the LP residual is deemphasized in the low SNR regions, the LP filter for each of the 30 ms segment dominates the system characteristics in the reconstructed speech signal. Therefore it is necessary to improve the system level spectral characteristics. One way of doing this is to obtain the LPCs for shorter segments from noisy speech, so that for the high SNR segments the LP filter will be close to the true one. For the other segments the filter characteristics are deemphasized in the reconstruction due to deemphasis of the corresponding residual. But unfortunately, we do not have a good method of estimating the LP filter for short (< 10 ms) segments.

One way to address this problem is to use a low (2-3) order inverse filter for the noisy speech first, and then use a high (8-10) order LP analysis on the residual from the low order inverse filter using short (5 ms) segments, with a shift of 2 ms between segments. Now the LP filter is a cascade of the two filters for each 2 ms intervals. The residual is computed by using the two levels of inverse filtering, and it is manipulated as described before. The modified residual is used to excite the time varying cascaded filter updated every 2 ms to generate the enhanced speech.

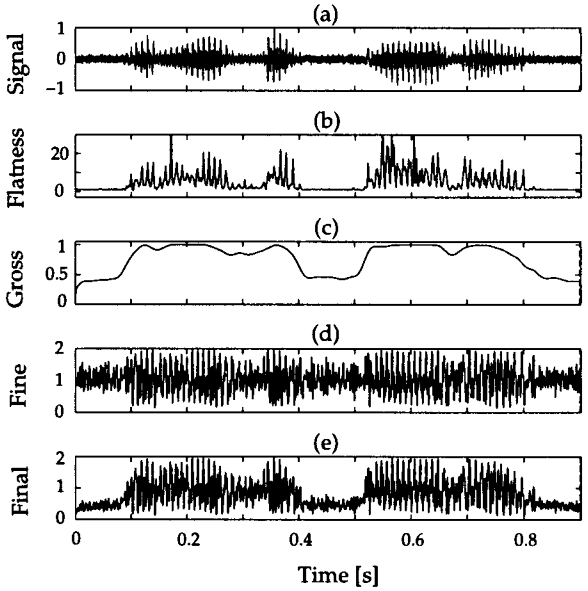

Figure 1: (a) Speech signal for the utterance "any dictionary" with an average SNR of 10 dB. (b) The ratio of energy values for 10 dB SNR case for each 2 ms frame. (c) The smoothed gross weighting function. (d) The fine weighting function. (e) The final weighting function.

4. STUDIES ON DIFFERENT TYPES OF NOISES

The proposed method works fairly well for different noise levels. The degradation is gradual and graceful as the noise level is increased. It is important to note that the thresholds for deriving the weighting function could be adjusted so as to obtain an acceptable trade off between reduction in noise annoyance and degradation in speech quality based on perceptual impressions of the enhanced speech.

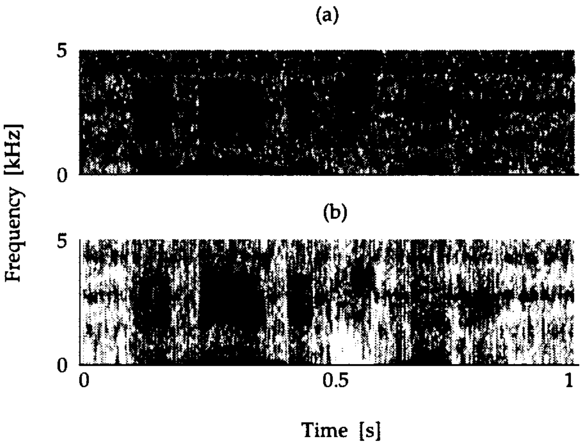

Figure 2: (a) Spectrogram for 10 dB SNR speech corrupted by cockpit noise. (b) Spectrogram for enhanced speech using spectral level manipulation besides gross and fine level weighting.

We tested our algorithm on several samples of speech corrupted by AWGN. Informal listening tests indicated considerable reduction of annoyance in the processed samples. The method works well even for colored additive noise. Fig. 2(a) shows the spectrogram of speech corrupted by noise recorded in the cockpit of an F16 aircraft [11]. The average SNR is adjusted to 10 dB. We note from the spectrogram that the cockpit noise exhibits both wideband as well as narrowband (spectral lines at approximately 3000 and 4500 Hz) nature. Fig. 2(b) shows the spectrogram of enhanced speech. The enhancement is carried out using the algorithm proposed in the previous section in three iterations. We found that the enhancement was better when carried out in small steps over two or three iterations rather than in one large step. In each iteration, mild enhancement can be obtained by using suitable values for the thresholds in the *tanh*(x) mapping function. For nonwhite noise situations, the minimum value of the inverse flatness is used to derive suitable values for the thresholds in the mapping function.

5. SUMMARY AND CONCLUSIONS

In this paper we have presented a new approach for enhancement of speech based on LP residual. The method uses the fact that in noisy speech the SNR is a function of time and also of frequency. By enhancing the high SNR regions relative to the low SNR regions, the annoyance due to the background noise is reduced without significantly distorting the quality of speech. This is accomplished by identifying the low and high SNR regions based on inverse spectral flatness characteristics in short (2 ms) of time frames. The inverse spectral flatness information is derived using the ratio of energies in the LP residual of the speech and the noisy signal in each 2 ms segments. The spectral flatness characteristics are used to derive a weighting function for the residual signal at gross level, and the residual signal energy to derive the weighting function at finer level. The two weighting functions are multiplied to get the overall weighting function for the residual. Since no direct spectral manipulation is involved, this method does not produce the type of distortions which the spectral subtraction and smoothing methods produce. It is interesting to note that the various parameter values used in processing, such as LP order, analysis frame sizes etc., are not critical.

6. REFERENCES

1. J. Junqua and J. Haton, *Robustness in Automatic Speech Recognition: Fundamentals and Applications*, Kluwer Academic, Boston, 1996.
2. S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.
3. Y. M. Cheng and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence," *IEEE Trans. Signal Processing*, vol. 39, no. 9, pp. 1943–1954, Sep. 1991.
4. Y. Ephraim and H. L. VanTrees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, Jan. 1994.
5. K. Y. Lee and K. Shirai, "Efficient recursive estimation for speech enhancement in colored noise," *IEEE Signal Processing Lett.*, vol. 3, no. 7, pp. 196–199, Jul. 1996.
6. C. Avendano, H. Hermansky, M. Vis and A. Bayya, "Adaptive speech enhancement using frequency-specific SNR estimates," *Proc. III IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*, Basking Ridge, New Jersey, pp. 65–68, Sep. 1996.
7. T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 4, pp. 309–319, Aug. 1979.
8. B. Yegnanarayana and P. Satyanarayana Murthy, "Source–System windowing for speech analysis and synthesis," *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 2, pp. 133–137, Mar. 1996.
9. S. J. Leon, *Linear Algebra with Applications*, Macmillan, New York, 1990.
10. P. Satyanarayana Murthy and B. Yegnanarayana, "Robustness of group delay based method for extraction of significant instants of excitation from speech signals," submitted to *IEEE Trans. Speech and Audio Processing.*
11. *IEEE Signal Processing Information Base*, Web site – http://spib.rice.edu/spib/select_noise.html