ENHANCEMENT OF REVERBERANT SPEECH USING LP RESIDUAL

B. Yegnanarayana
Dept. of CSE, IIT, Madras 600 036, India
C. Avendano
Dept. of EE, OGI, Portland, OR 97291, U.S.A.

P. Satyanarayana Murthy
Dept. of EE, IIT, Madras, India
H. Hermansky
Dept. of EE, OGI, Portland, OR, U.S.A.

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 in three different regions of the speech signal, namely, high SRR region, low SRR region and only reverberation component region. A weighting function is derived to modify the LP residual. The weighted residual samples are used to excite the time–varying LP all–pole filter to obtain perceptually enhanced speech.

1. 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 a speakerphone 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.

Several methods have been proposed for enhancement of speech degraded by reverberation [1–3]. Normally degraded (additive noise or reverberant) speech is processed assuming that the degradation has long term stationary characteristics relative to speech. 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 [2]. The reverberant speech is passed through an inverse filter for the room response to dereverberate the 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 Reverberant component Ratio (SRR) 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.

There appears to be a need to look at the problem of enhancement of reverberant speech with more focus 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. In such an attempt there will be more focus on speech than on the degradation in the process of enhancement. This point of view is also reasonable, since speech is a nonstationary signal, with signal energy varying over a wide ($\approx$ 60 dB) dynamic range both in temporal and spectral domains. Therefore the signal to degradation ratio will be varying over short (10–30 ms) segments of data.

Since dereverberation is not a realizable task, the focus should be on enhancement, but not 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. The levels of the higher reverberation regions, if identified, could be reduced. In the regions where there is only reverberant component, such as silence regions, the levels could be reduced to very low values. Perception of the overall speech is significantly influenced by the high signal energy regions, thus giving an impression of enhancement of degraded speech. Thus the criterion for improvement is not based on all speech segments, but only on high direct path signal component regions.

The method proposed in this paper 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 10–30 ms frames used in speech analysis based on quasistationary assumption. In Section 2 we discuss the model of reverberant speech and some of its characteristics. In Section 3 the steps for processing degraded speech are discussed. In particular, the importance of processing the linear prediction (LP) residual signal is emphasized, since most of the conventional ap-

0-7803-4428-6/98 $10.00 © 1998 IEEE

proaches tend to ignore the details of the residual signal. We also present some experimental results in this section.

2. CHARACTERISTICS OF REVERBERANT SPEECH

In this section we will examine the characteristics of reverberant speech to determine clues for processing the speech for enhancement.

The effects of reverberation can be seen by comparing the signal waveforms for clean and reverberant speech signals shown in Fig. 1. The clean speech has clear damped sinusoidal–like pattern within each pitch cycle, whereas the reverberant speech is smeared within each cycle (region AB). The smearing of signal within each pitch cycle is more prominent when the gross envelope of the signal waveform is decaying as in the region BC in the figure. The smearing extends for several pitch cycles due to the influence of the large amplitude signal component in the region AB. The reverberation tail component only is present in the low amplitude/silence regions such as region CD in the figure.

Some features of reverberation effects can be seen more clearly in the LP residual waveform. The residual signal is computed for a segment of 2 ms at every sampling instant, using a 5th order autocorrelation LP analysis. The residual signal for reverberant speech signal has 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 amplitude at the epochs (instants of significant excitation i.e. instants of glottal closure) is high relative to the signal in the rest of the pitch cycle, as 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 decaying nature of the overall signal amplitudes, the reverberation effects of the preceding speech dominates over the direct component. In the region CD, the residual signal is mainly due to the reverberation. Whenever the direct component of speech is higher than the reverberant component, the LP residual signal at the epochs is well behaved with significant energy around the instants of glottal closure. It is such regions that we need to identify, so that the signals in those regions can be processed to enhance the direct component over the reverberant component. The signal 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 over the signal in the rest of the pitch cycle.

First of all it is necessary to identify these three different regions in reverberant speech. For this purpose the normalized error (η) of clean and reverberant speech is computed at every sampling instant using a 5th order autocorrelation LP analysis on a frame of size 2 ms. The normalized errors for both clean and reverberant speech appear similar in the high SRR regions. But overall the normalized error for reverberant speech is lower than that for the clean speech. Note that η also gives a measure of spectral flatness [4].

A closer examination of the normalized error plot reveals that within each pitch cycle the error is maximum just before the region of glottal closure. This is because the residual signal amplitude is high in this region. These points of maximum η within each pitch cycle can easily be identified in the high SRR regions such as AB. It is more difficult to see this distinction between open and closed glottis regions in the low SRR regions such as BC. The normalized error in the purely reverberant region as in CD show lower values, but no features such as periodic peaks.

The important point to be noted is that the enhancement needs to be done differently in different segments due to variation of short–time characteristics of speech in temporal and spectral domains [5].

3. PROCESSING REVERBERANT SPEECH USING LP RESIDUAL FOR ENHANCEMENT

For processing reverberant speech for enhancement we propose to manipulate the LP residual signal in short (2 ms) and in long (> 20 ms) segments in a selective manner. The manipulation basically involves weighting the residual samples appropriately. Manipulation of the residual signal is more appropriate than the speech signal, especially for short (2 ms) segments, as the residual samples are nearly uncorrelated. On the other hand, in the manipulation of the speech signal directly, the choice of window size and shape will significantly affect the performance. It is interesting to note that after manipulation, any distortion in the processed residual signal are smoothed out by the all–pole filter used for synthesis.

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

As mentioned earlier, processing of the LP residual involves determination of appropriate weight function for the residual. The weight function is derived for manipulating the residual both at the fine (within pitch cycle) level and at the gross level. To derive the weight function, we need to identify the different SRR regions both at fine and at gross levels from the reverberant speech signal. That is, we need to determine the three types of regions such as AB, BC and CD in Fig. 1 and also the regions around the instants of glottal closure in AB. The regions are identified using the properties of the LP residual signal for reverberant speech. The regions at the gross level are determined using the statistics of the LP residual signal. In high SRR regions the entropy of the distribution of the LP residual samples is low compared to the entropy in the low SRR regions. This is because the LP residual samples exhibit a Gaussian like probability den-

sity function (pdf) 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 due to strong excitations of the vocal tract, skew the pdf and so the entropy is low. To compute the entropy, the pdf of the samples in 20 ms segments of the LP residual is estimated. The use of a longer (20 ms) segment is to obtain a good estimate of the histograms of the samples and hence their pdf. The entropy H_k for the frame k is computed using the following expression [6] :

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

where p_i is the probability estimated in the ith bin of the histogram and M is the number of bins in the histogram. The number of bins (M) is chosen to be 7, though this value is not critical and can be any value between 5 and 20, so that there are enough LP residual samples per bin. The entropy is computed using 20 ms frames for every 10 ms. The result is shown in Fig. 2(c). The smoothed entropy function (Fig. 2(d)) is derived by repeating each entropy value 80 times (10 ms at 8 kHz sampling rate) and then smoothing this interpolated function with a 600–point mean smoothing filter. From the smoothed entropy function the gross weighting function is derived using a nonlinear mapping function (of $tanh$ type) between smoothed entropy and weighting function value, such that large entropy values get mapped to low weights and vice versa. The setting of various thresholds in the mapping function is not critical most of the time.

From the gross weighting function the three different types of SRR regions can be identified. The regions of rising and high values of weight function correspond to the high SRR regions corresponding to AB in Fig. 1. The falling portion corresponds to the low SRR region corresponding to BC in Fig. 1. The low weight function regions correspond to the reverberant component regions such as CD in Fig. 1. The normalised error (η) computed at each sample for a frame of 2 ms using a 5th order LP analysis is shown in Fig. 2(f). The normalised error provides relative weighting of short segments within a pitch cycle in the high SRR regions. The overall weight function is obtained by multiplying the gross weighting function with the normalized error. The resulting weight function, shown in Fig. 2(g), is used to derive a modified residual signal. The modified residual signal is used to excite the 5th order all–pole filter. The filter is updated at every sampling instant.

The performance of the proposed method is illustrated on speech data spoken by a female speaker collected under reverberant conditions. The speech data was collected in a normal office room with the microphone placed about 5′ away from the speaker (see Fig. 2(b)). Speech data was also collected simultaneously close to the speaker to use it as a clean speech signal (see Fig. 2(a)) for comparison. The differenced reverberant speech signal data was processed using the algorithm presented above. The signal waveform and its spectrogram are given in Fig. 3 for clean speech signal, reverberant speech signal, and the processed speech signal. From the spectrograms it is evident that the effects of reverberation are significantly reduced. Perceptually also the processed signal sounds less reverberant than the unprocessed one. The results show that the values of thresholds used in deriving the weight function are not very critical. They provide a tradeoff between quality and enhancement in the processed signal.

4. 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 even 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 pitch cycle) level one can achieve enhancement of reverberant speech. The processing was done by weighting the LP residual. The weighting 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.

5. REFERENCES

[1] in *Speech Enhancement* (J. S. Lim, ed.), New Jersey: Prentice Hall, 1983.

[2] 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.

[3] C. Avendaño and H. Hermansky, "Study on the dereverberation of speech based on temporal envelope filtering," in *Proceedings of Int. Conf. Spoken Language Processing*, (Philadelphia), pp. 889–892, Oct. 1996.

[4] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[5] B. Yegnanarayana, C. Avendaño, H. Hermansky, and P. Satyanarayana Murthy, "Processing linear prediction residual for speech enhancement," in *Proceedings of EUROSPEECH'97*, (Patras, Greece), pp. 1399–1402, Sept. 1997.

[6] J. G. Proakis, *Digital Communications*. Singapore: McGraw–Hill, 1989.

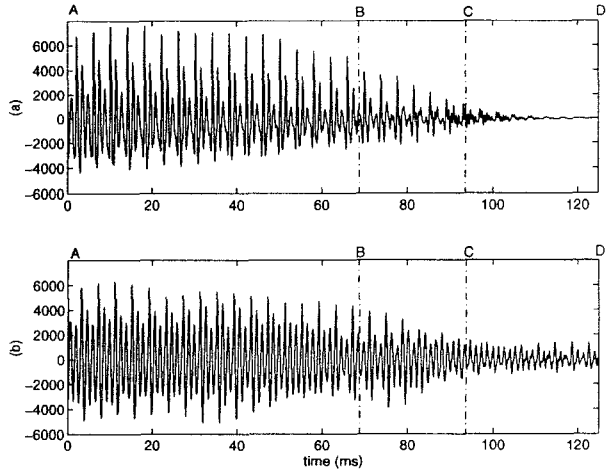

Figure 1: (a) Clean speech signal. (b) Reverberant speech signal.

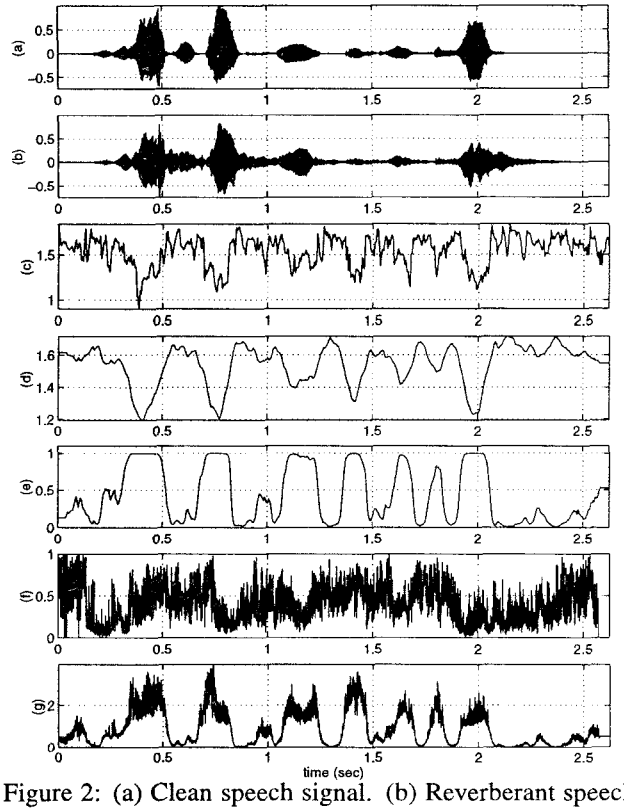

Figure 2: (a) Clean speech signal. (b) Reverberant speech signal. (c) The moving entropy computed from the LP residual of reverberant speech in (b). (d) Interpolated and smoothed version of the entropy function in (c). (e) Gross weight function. (f) Normalised prediction error. (g) Overall weight function.

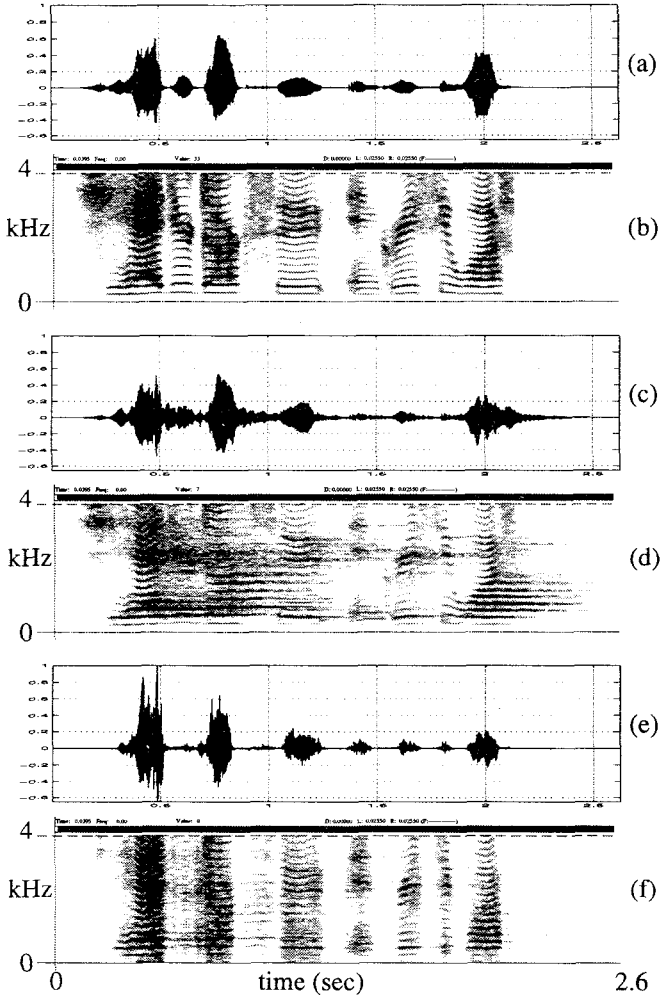

Figure 3: (a) Clean speech. (b) Spectrogram for clean speech. (c) Speech degraded by reverberation. (d) Spectrogram for speech degraded by reverberation. (e) Enhanced speech using gross and fine level weighting of the LP residual. (f) Spectrogram for enhanced speech using gross and fine level weighting of the residual.