Enhancement of Reverberant Speech using Excitation Source Information

M. Chaitanya, S. R. Mahadeva Prasanna and B. Yegnanarayana

Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600 036, INDIA
Email: {chaitanya, prasanna, yegna}@cs.iitm.ernet.in

Abstract

This paper proposes a method for the enhancement of reverberant speech using the knowledge of the excitation source of speech production. The degradation level in the reverberant speech is measured in terms of Speech-to-Reverberation component Ratio (SRR). From perception and processing point of view high SRR regions are important. Hence the proposed method identifies and enhances the speech in high SRR regions. The high SRR regions are identified using the Hilbert envelope of the Linear Prediction (LP) residual, which contains information about the excitation source of speech production. The Hilbert envelope of the LP residual derived from the reverberant speech is processed by the covariance analysis to derive the weight function. The LP residual of the reverberant speech is multiplied with the weight function to enhance the excitations of speech in the high SRR regions. The speech signal synthesized from the modified LP residual is found to be less reverberant.

1. Introduction

The speech collected from a reverberant environment not only contains the direct component of speech but also the component due to reflections from various surfaces (reverberation). The speech may also contain component due to background noise. However, the scope of the present study is to enhance the speech mainly against reverberation. The level of degradation in the reverberant speech can be measured in terms of Signal-to-Reverberation component Ratio (SRR) [1]. Depending on the level of reflected component, the reverberant speech can be classified into high SRR, low SRR, and only reverberation regions [1]. From the perception and processing point of view high SRR regions are important [1]. Thus the reverberation effect may be minimized by emphasizing the direct component of speech in the high SRR regions. For this first the high SRR regions needs to be identified. In the present work the high SRR regions are identified using the excitation source information.

Reverberant speech signal can be expressed as [1]

$$s(n) = y(n) + \sum_{k=1}^{N} b_k y(n-k) \tag{1}$$

where $\{s(n)\}$ is the reverberant signal, $\{y(n)\}$ is the direct component of speech signal and $\{b_k\}$ correspond to the filter coefficients for the impulse response of the room. As it is given in Eqn. (1), the degrading component in the reverberant signal is also like speech. Therefore the challenge is to differentiate between the direct and reflected components of speech and then emphasizing the direct component. As it is difficult to separate the direct and reflected components of speech, we try to emphasize the direct component by emphasizing the excitation of speech in the high SRR regions. The emphasis of the excitation in high SRR regions reduces the reverberation effect in the processed signal [1].

The speech from the acoustical environment can be collected over a single or a set of spatially distributed microphones. Accordingly, we can classify the enhancement techniques into single and multichannel cases [2–4]. In multichannel case the high SRR regions can be identified by using signals from multiple microphones. However, it is difficult to identify the high SRR regions in the single channel case. A method has been proposed for processing reverberant speech by analyzing the speech signal in small segments (1-3 ms) [1]. In this paper, we propose a method for processing reverberant speech in single channel case using the excitation source information.

The basis for the proposed method may be explained as follows: The excitation of speech (especially voiced regions) can be treated to a first approximation as a sequence of impulses, which may be termed as speech epochs. Similarly, the excitation of reverberant speech will also be a sequence of impulses, which may be termed as reflected epochs. The level of reverberation in the degraded speech depends on the relative strengths of speech and reflected epochs. The reverberation effect may be minimized by emphasizing the speech epochs relative to the reflected epochs and synthesizing speech from the modified excitation signal.

This paper is organized as follows: Section 2 discusses about the extraction of the excitation source information. The proposed method for enhancement of speech is discussed in Section 3. The experimental results are given in Section 4. The summary of the present work and the scope for future work are given in Section 5.

2. Excitation Source Information

Speech is the result of convolution of the excitation sequence with the vocal tract system features. The excitation source information can be extracted from the speech signal by the Linear Prediction (LP) analysis [5]. The LP residual is the error between the actual and the predicted sequence which is given by

$$e(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \tag{2}$$

where p is the order of prediction and $\{a_k\}$ are the Linear Prediction Coefficients (LPC) obtained by LP analysis. Since $\{a_k\}$ models the vocal tract system features, the LP residual $e(n)$ mostly contains information about the excitation source.

A segment of clean speech signal and its LP residual are shown in Figs. 1(a) and (b), respectively. Whenever there is significant excitation of the vocal tract system it is indicated by a large error in the LP residual. This can be seen well in case of voiced speech, where the significant excitation within a pitch period coincides with the Glottal Cloure (GC) instant. GC instant is the instant at which closure of vocal folds takes place within a pitch period. Eventhough the LP residual mostly contains the excitation source information, there are difficulties in using it directly for further processing. This is due to the phase of the residual which results in signal of either polarity around the instants of significant excitation. The effect of phase can be minimized using the Hilbert envelope of the LP residual which is defined as [6]

$$h_e(n) = \sqrt{e^2(n) + e_h^2(n)} \tag{3}$$

where, $e_h(n)$ is the Hilbert transform of $e(n)$ and is computed as

$$e_h(n) = \begin{cases} IDFT[-jE(\omega)], & 0 < \omega < \pi \\ IDFT[jE(\omega)], & 0 > \omega > -\pi \\ 0, & \omega = 0, \pi \end{cases}$$

where, IDFT is the Inverse Discrete Fourier Transform and $E(w)$ is the discrete Fourier transform of $e(n)$. The Hilbert envelope of the LP residual is shown in Fig. 1(c). The representation of the excitation source information, mainly, the instants of significant excitation is better in the Hilbert envelope of the LP residual. Hence, Hilbert envelope of the LP residual is used as the excitation source information in this study.

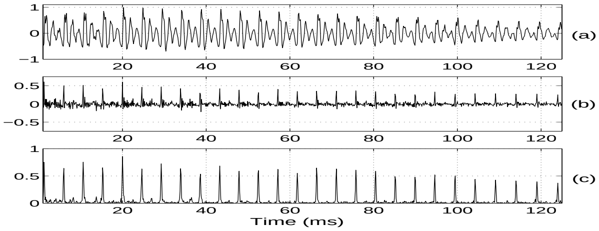

Figure 1: (a) Segment of speech signal and its (b) LP residual, and (c) Hilbert envelope of the LP residual.

3. Enhancement of Reverberant Speech

Hilbert envelopes of the LP residuals of clean speech and the corresponding reverberant speech are shown in Figs. 2 (a) and (b), respectively. It can be observed that in the Hilbert envelope of reverberant speech, in addition to the speech epochs, there are reflected epochs due to the reflections from the surfaces. However, the reflected epochs occur at random instants. In order to distinguish between the speech epochs and the reflected epochs, covariance analysis is performed on the Hilbert envelope of the LP residual. For every frame of N samples of the Hilbert envelope of the LP residual, the covariance sequence $\{\varphi_k\}$ is computed as

$$\varphi_k = \sum_{n=1}^{N} h_e(n) h_e(n+k) \qquad k = 1, 2, \ldots, N \tag{4}$$

The covariance sequences are computed for the Hilbert envelope of the LP residual by considering frames of 20 ms with a frame shift of 2 ms. The obtained covariance sequences are time aligned with the corresponding Hilbert envelope frames by the cross-correlation approach [7]. These time aligned sequences are added to get probable regions of the speech epochs. A segment of the Hilbert envelope is shown in Fig. 3(a) and the time aligned covariance sequences are shown in Figs. 3(b)-(f). The coherently-added covariance signal is shown in Fig. 3(g). It can be observed that in the coherently-added covariance signal the speech epochs are added coherently, where as the reflected epochs are spread out in time.

The coherently-added covariance signal can be used for detecting the high SRR regions in the reverberant speech. Typically, the voiced regions will be high SRR in nature. Most of the speech produced is of voiced type and also due to coarticulation effect the information about the

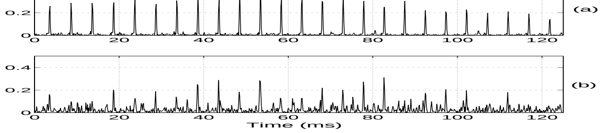

Figure 2: Hilbert envelope of the LP residual of (a) clean speech and the corresponding (b) reverberant speech.

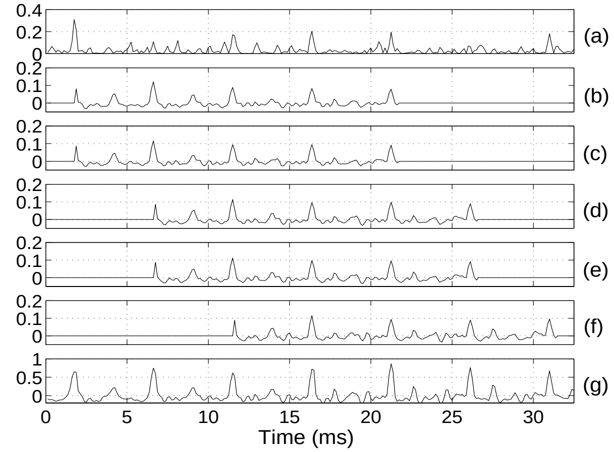

Figure 3: (a) Segment of Hilbert envelope of the LP residual, (b)-(f) covariance sequences of five consecutive frames and (g) coherently-added covariance signal

consonants (unvoiced type) will also be reflected in the voiced regions. Thus, identifying and emphasizing the speech epochs in the voiced regions increases the level of direct component and hence the synthesized speech will be less reverberant. The excitation will be mostly periodic in nature in case of voiced speech and this periodicity is not observed in unvoiced and nonspeech regions. The periodicity in the coherently-added covariance signal may be identified by the autocorrelation analysis. For illustration a 30 ms frame of the coherently-added covariance signal computed from the voiced segment of reverberant speech and its autocorrelation sequence are shown in Figs. 4(a) and (b), respectively. The strength of the first peak (after the center peak) in the autocorrelation sequence is an indication of the voicing level in the segment, which is high in this case. Similarly autocorrelation analysis performed for a 30 ms frame of nonspeech frame (see Fig. 4(c)) is shown in Fig. 4(d). As expected the strength of the first peak is low in this case. Thus the first peak strength (Ps) obtained by performing autocorrelation analysis on the frames of the coherently-added covariance signal for every sample shift gives an indication of the measure of the sample belonging to the voiced speech region. Based on this feature a gross weight function can be derived which may be used to emphasize the voiced regions (high SRR) in the degraded speech signal.

Further emphasis of high SRR regions within the identified voiced regions can be done as follows: The direct component of speech will be relatively high around the speech epochs as the excitation of the vocal tract system is high at these regions. Thus, these regions correspond to high SRR and hence may be emphasized to reduce the reverberation effect in the speech. The coherently-added covariance signal of the voiced regions can be further processed to obtain a fine weight function which gives emphasis for the high SRR regions. The energy of the coherently-added covariance signal is calculated over a frame size of 5 ms and frame shift of one sample and is normalised with its running mean. The obtained values are mapped using a nonlinear mapping function to obtain a fine weight function. A segment of Hilbert envelope of the LP residual for clean and corresponding reverberant speech are shown in Figs. 5(a) and (b), respectively. The coherently-added covariance signal and corresponding fine weight function are shown in Figs. 5(c) and (d), respectively. It can be observed that the regions corresponding to high SRR are given higher weightage relative to low SRR regions.

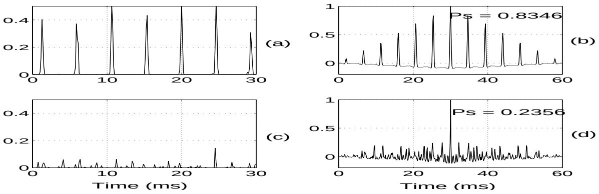

Figure 4: (a) A 30 ms of voiced speech frame and its (b) autocorrelation sequence. (c) A 30 ms nonspeech frame and its (d) autocorrelation sequence.

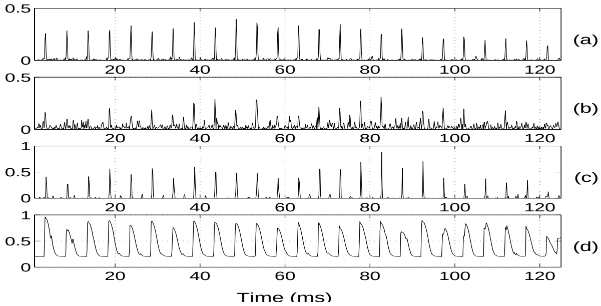

Figure 5: Segment of Hilbert envelope of the LP residual of (a) clean speech, corresponding (b) reverberant speech, (c) coherently-added covariance signal and (d) fine weight function.

The LP residual of the reverberant speech is modified using the gross and fine weight functions. As the samples in the LP residual signal are less correlated compared to the samples in the speech signal, modifying the LP residual may introduce less distortion in the processed signal. The enhanced speech signal is synthesized by exciting the time-varying filter using the modified LP residual. The parameters of the time-varying filter are derived from the reverberant speech.

4. Experimental Results

A reverberant speech signal is shown in Fig. 6(a). Hilbert envelope of the LP residual derived from the reverberant speech by performing 10^{th} order LP analysis is shown in Fig. 6(b). Covariance analysis is performed on the Hilbert envelope of the LP residual to obtain coherently-added covariance signal and is shown in Fig. 6(c). Autocorrelation analysis is performed on the coherently-added covariance signal to obtain the peak strength values. The peak strength values are used in deriving a gross weight function shown in Fig. 6(d). Using the coherently-added covariance signal, fine weight function is derived for speech regions as discussed earlier (see Fig. 6(e)). The LP residual modified using both gross and fine weight functions is shown in Fig. 6(f). The modified LP residual is used to synthesize the enhanced speech, which is shown in Fig. 6(g).

The narrowband spectrograms of clean, degraded and corresponding enhanced speech signals are shown in Fig. 7. Reverberant tails following the speech regions are attenuated and as the direct component of speech signal is emphasized, the enhanced speech looks similar to the clean speech. The degraded and the corresponding enhanced speech signals obtained by the proposed method are available for listening at
http://speech.cs.iitm.ernet.in/Main/result/onech.html.

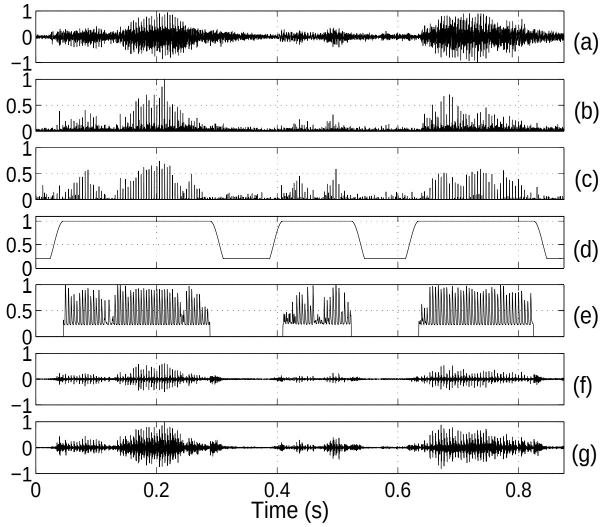

Figure 6: (a) Reverberant speech signal, and its (b) Hilbert envelope, (c) coherently-added covariance signal, (d) gross weight function, (e) fine weight function, (f) modified LP residual, (g) enhanced speech signal.

5. Conclusions

In this paper a method was proposed for enhancing speech degraded by reverberation. The proposed method exploits the knowledge of the excitation source of speech production to identify the high SRR regions. The high SRR regions are emphasized by giving higher weightage to the excitations of speech in the LP residual. The synthesized speech signal is found to be perceptually less reverberant compared to the degraded signal. The time-varying filter parameters currently used for synthesizing the enhanced speech signal are derived from the degraded signal. The performance may be improved by further enhancing time-varying filter parameters.

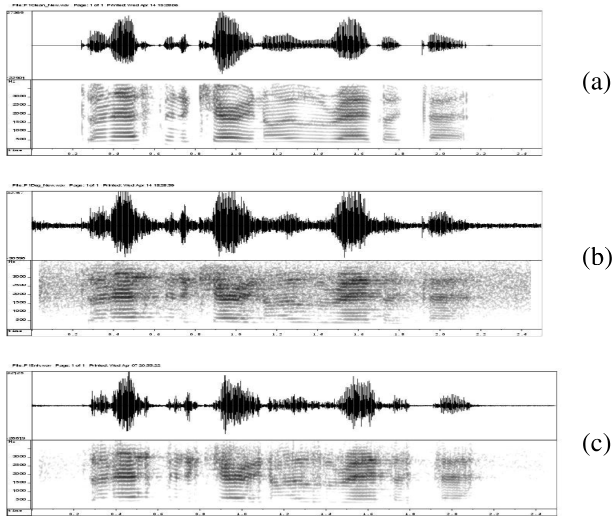

Figure 7: Speech signals and narrowband spectrograms of (a) clean, (b) degraded and (c) enhanced signals, respectively

6. References

[1] B. Yegnanarayana and P. S. Murthy, “Enhancement of reverberant speech using LP residual signal,” *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 267–281, May 2000.

[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] J. L. Flanagan, J. D. Jonston, R. Zahn, and G. W. Elko, “Computer-steered microphone arrays for sound transduction in large rooms,” *J. Acoust. Soc. Amer.*, vol. 78(5), pp. 1508–1518, 1985.

[4] D. V. Compernolle, “Switching adaptive filters for enhancing noisy and reveberant speech from microphone array recordings,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Albuquerque, New Mexico, USA), pp. 833–836, Apr. 1990.

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

[6] 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, pp. 309–319, Aug. 1979.

[7] B. Yegnanarayana, S. R. M. Prasanna, and K. S. Rao, “Speech enhancement using excitation source information,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. I, (Orlando, FL, USA), pp. 541–544, May 2002.