SPEECH ENHANCEMENT USING EXCITATION SOURCE INFORMATION

B. Yegnanarayana, S. R. Mahadeva Prasanna and K. Sreenivasa Rao

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

ABSTRACT

This paper proposes an approach for processing speech from multiple microphones to enhance speech degraded by noise and reverberation. The approach is based on exploiting the features of the excitation source in speech production. In particular, the characteristics of voiced speech can be used to derive a coherently added signal from the linear prediction (LP) residuals of the degraded speech data from different microphones. A weight function is derived from the coherently added signal. For coherent addition the time-delay between a pair of microphones is estimated using the knowledge of the source information present in the LP residual. The enhanced speech is generated by exciting the time varying all-pole filter with the weighted LP residual.

1. INTRODUCTION

Speech signal collected with a microphone placed in a live room is degraded by both additive noise and reverberation. The additive noise effect is generally independent of the relative positions of the speaker and the microphone. On the other hand, the reverberation effect is critically dependent on the position of the microphone and the speaker inside the room. The direct component of speech is reduced with increasing distance of the microphone from the speaker, and hence the direct signal to reverberant component ratio (SR-R) of speech decreases [1], [2]. The signal to noise ratio (SNR) due to additive noise also decreases with increasing distance of the microphone from the speaker, but this reduction can be compensated by increasing the volume of the source of speech. But the SRR is unaffected by the increase in volume.

The challenge in enhancement of speech is to raise the component of speech over the background noise and to reduce the effects of reverberation. The main problem in processing of speech for enhancement is the nonstationary nature of the speech production process. The temporal and spectral characteristics of speech change continuously, both in the nature and in the energy. Thus the SNR is both a function of time as well as a function of frequency. Therefore it is difficult to estimate the characteristics of the degraded speech signal. Thus methods based on spectral subtraction for noise reduction and deconvolution of room response for dereverberation may not be satisfactory [3], [4].

In this paper we propose a method for enhancing speech by focusing on the characteristics of speech production. In particular, voiced speech is produced as a result of excitation of the vocal tract system with quasi-periodic glottal pulses. The significant excitation in each glottal cycle takes place at the instant of glottal closure. The strength of the excitation of the vocal tract system depends on the rate at the instant of glottal closure. In fact it is this strength that enables us to perceive speech in spite of degradation in speech. For example, it is almost impossible to perceive the speech message in degraded whispered speech.

By locating the instant of glottal closure in each cycle, it is possible to enhance speech in the region of the instant relative to other regions. Due to high strength of excitation, the SNR of speech is high around these instants compared to other regions. Thus it is possible to enhance speech by exploiting the characteristics of excitation source in speech production. Linear prediction analysis can be used to derive the source characteristics [5]. Short (1-3 ms) segment analysis of the linear prediction residual was used to develop methods for enhancement of speech collected by a single microphone [1]. The method was based on deriving a weight function for the LP residual using the properties of the residual for different degradations. If speech data is available from two or three spatially distributed microphones, then it is possible to derive a weight function for the LP residual to reduce the effects of additive noise and reverberation in the enhanced speech. This paper describes a new approach for enhancement of speech from multiple microphones. We show that the effect of reverberation can be reduced by coherently combining the source information derived from the microphone outputs. In Section 2 we describe the proposed method for speech enhancement, which includes estimation of time-delay, coherent addition of source information, derivation of weight function, modification of the LP residual and synthesis of enhanced

0-7803-7402-9/02/$17.00©2002 IEEE

speech. Experimental results are presented in Section 3 in the form of signal waveforms and the corresponding speech spectrograms. In the final section some issues for further study are discussed.

2. SPEECH ENHANCEMENT USING MODIFIED LP RESIDUAL

The basis for the proposed enhancement method is the fact that in voiced speech the significant excitation of the vocal tract system takes place at the instant of glottal closure [6, 7, 8, 9, 10]. The excitation is significant because the strength of excitation is highest at that instant in each glottal cycle. The SNR of the speech signal is also high in the region around the instant which can be exploited for speech enhancement. The signals from multiple microphones have the same sequence of significant instants, except for the fixed delay between a pair of microphones. For speech degraded due to reverberation, there will be many other significant instants due to reflections from the surfaces. But these instants will occur at random instants at each microphone. Therefore, if the microphone outputs are added after compensating for the time-dealys, then the signal gets added coherently at the significant instants corresponding to the direct path. The reverberant components get added incoherently. While there will be improvement in the SNR due to this coherent addition, the reverberant component will still be present.

One possible solution to improve SRR and simultaneously increase SNR is to suitably modify the LP residual signal. The idea is to generate a weight function for the LP residual, which enhances the coherent part around the significant instants relative to the other regions. Since the LP residual has both positive and negative samples depending on the phase, the strength of the LP residual at each instant is obtained by computing the Hilbert envelope of the LP residual signal [7, 11]. The Hilbert envelope $\hat{e}(n)$ of the residual signal $e(n)$ is obtained as

$$\hat{e}(n) = \sqrt{e^2(n) + e_H^2(n)} \tag{1}$$

where $e_H(n)$ is the Hilbert transform of $e(n)$.

The Hilbert transform of a signal $e(n)$ is obtained by exchanging the real and imaginary parts of the DFT of $e(n)$, and then computing the IDFT. The amplitude of the Hilbert envelope gives an indication of the strength of excitation at that instant, and the amplitudes are typically large around the instants of glottal closure. Fig.1 shows the LP residual and the Hilbert envelope of a 10^{th} order LP residual of a segment of clean voiced speech.

Due to the effects of noise and reverberation, there will be several large amplitude spikes in the Hilbert envelope of the LP residual. To reduce the effects of these spikes, one can determine the Hilbert envelopes of the LP residuals from several microphones, and add them coherently. Fig.2 shows the Hilbert envelopes of the LP residuals from three spatially distributed microphones.

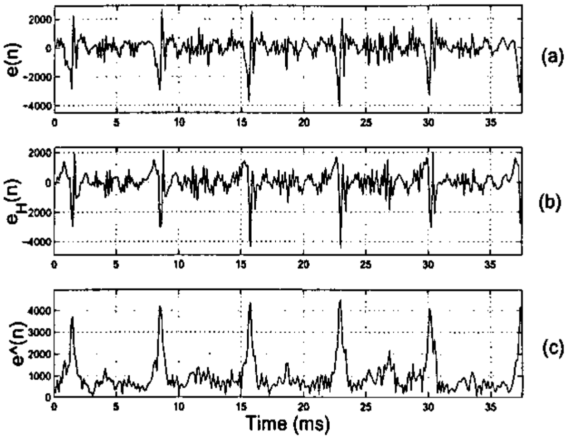

Fig. 1. (a) LP residual of a voiced segment ($e(n)$), (b) Hilbert transform ($e_H(n)$) of the LP residual, and (c) Hilbert envelope ($\hat{e}(n)$).

For coherent addition the time-delay of the signal between two microphones need to be estimated. Crosscorrelation of a segment of $\hat{e}(n)$ with the corresponding segment from the second microphone gives a peak at the instant corresponding to the time-delay. The time delays are estimated for the two pairs of the three microphones. Let τ_{12} is the delay between microphones 1 and 2, and τ_{13} is the delay between microphones 1 and 3. The delay compensated Hilbert envelopes of microphones 2 and 3 are added to the first one. For this coherent addition the squares of the Hilbert envelopes are considered, and the resulting Hilbert envelope is the square root of this sum.

$$\hat{e}_c(n) = \sqrt{\hat{e}_1^2(n) + \hat{e}_2^2(n+\tau_{12}) + \hat{e}_3^2(n+\tau_{13})} \tag{2}$$

where $\hat{e}_c(n)$ is the coherently added Hilbert envelope, and $\hat{e}_1(n)$, $\hat{e}_2(n)$ and $\hat{e}_3(n)$ are the Hilbert envelopes of the LP residuals for microphones 1, 2 and 3, respectively. Fig.2(d) shows the coherently added Hilbert envelope and Fig.2(e) shows the incoherently added Hilbert envelope. For incoherent addition, τ_{12}=τ_{13}=0. It is evident that the strengths around the instants of glottal closure are emphasized relative to the strengths at other regions, thus reducing the effects of reverberation.

The nature of the coherently added Hilbert envelope is exploited to weight the residual. Weighting of the LP resid-

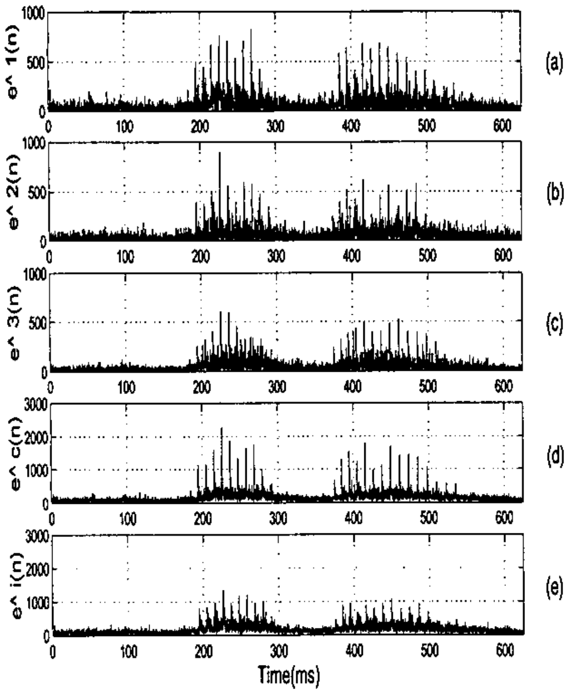

Fig. 2. (a), (b) and (c) are Hilbert envelopes of LP residuals of speech from three microphones ($\hat{e}_1(n)$, $\hat{e}_2(n)$, and $\hat{e}_3(n)$). (d) Hilbert envelope of the coherently added envelopes ($\hat{e}_c(n)$). (c) Hilbert envelope of the incoherently added envelopes ($\hat{e}_i(n)$).

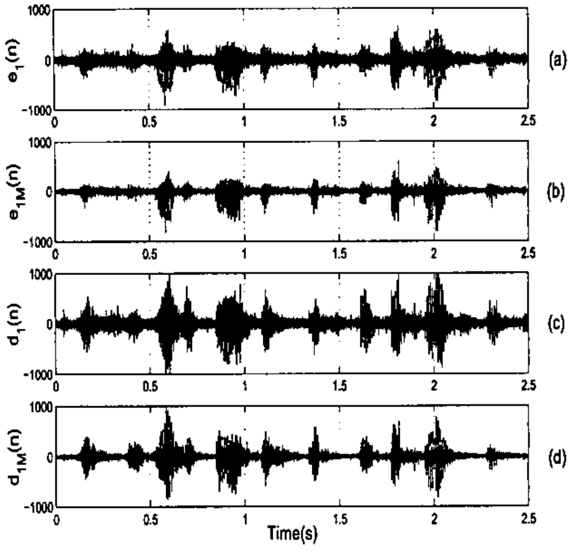

Fig. 3. (a) LP residual ($e_1(n)$), (b) Modified LP residual ($e_{1M}(n)$), (c) Degraded Speech ($d_1(n)$), and (d) Enhanced Speech ($d_{1M}(n)$).

ual $e_1(n)$ is done using the relation,

$$e_{1M}(n) = \frac{\sum_n e_1(n)\hat{e}_c(n)}{\sum_n \hat{e}_c(n)} \qquad (3)$$

The effect of this weighting is to emphasize the residual around the instants and smoothing the residual at other places.

Fig.3(a) and (b) shows the LP residual $e_1(n)$ and the modified LP residual $e_{1M}(n)$ respectively. This residual is used to excite the time varying all-pole filter to obtain the enhanced speech.

3. EXPERIMENTAL RESULTS

Speech signals are collected in a live room simultaneously from three spatially located microphones. The data was sampled at 8 kHz and stored as 16 bit integers. A 10^{th} order LP analysis is performed to derive the Hilbert envelopes of the LP residual for all the three microphone outputs. The time delays between pairs of microphone outputs are estimated using crosscorrelation of 100 ms segments of the Hilbert envelopes. The coherently added Hilbert envelope is used to derive a modified LP residual for the output signal from microphone 1. The enhanced speech is obtained by exciting the time varying 10^{th} order all-pole filter with the modified LP residual. The degraded speech and the corresponding enhanced speech by the proposed method are shown in Fig.3(c) and (d) respectively. The spectrograms of the degraded signal, enhanced signal and the coherently added speech signal are shown in Fig.4. The spectrogram clearly demonstrates the enhancement obtained when the the LP residual is weighted using the coherently added Hilbert envelope, compared to the case when the speech signals are coherently added.

The degraded speech, the output of the proposed enhancement method and the coherently added speech signal are available at the site:
http://speech.cs.iitm.ernet.in/Results/Enhance.html.

4. CONCLUSIONS

In this paper we have proposed a new method for enhancing speech collected from spatially distributed microphones. The enhancement technique is based on the properties of the excitation source for voiced speech. The most important property is that in voiced excitation the strength of excitation is maximum around the instant of glottal closure. We have used LP residual to derive the source features. The Hilbert envelope of a signal was used to derive the information of the strength of excitation. A weight function

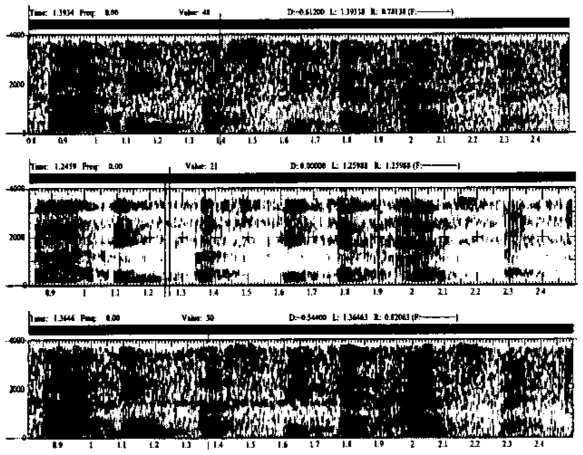

Fig. 4. Speech spectrograms of degraded signal (top), the enhanced signal (middle), and the coherently added speech signal (bottom).

was derived by combining coherently the delay compensated Hilbert envelopes of the LP residuals from the different microphones. The enhanced speech was derived by exciting the time varying all-pole filter with LP residual modified by the weight function.

The enhanced speech is significantly better compared to the coherently added speech signal, in the sense that the reverberation effects are reduced significantly. However, there is still significant degradation present in the enhanced speech may be due to noise and also due to poor representation of the all-pole filter. The all-pole filter was derived from the degraded speech. Further refinement may be possible by reducing the signal amplitudes in the region away from the instant of glottal closure in each glottal cycle. Also better all-pole filter can be derived from the degraded speech by processing the LP analysis around the instants of glottal closure.

5. REFERENCES

[1] B. Yegnanarayana and P. Satyanarayana Murthy, "Enhancement of reverberant speech using LP reisdual signal," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 267–281, May 2000.

[2] P. Satyanarayana, "Short segment analysis of speech for enhancement," *Ph.D. Thesis Dept. of Electrical Engineering IIT Madras, Chennai, India*, Feb. 1999.

[3] 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.

[4] 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.

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

[6] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction of voiced speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 562–570, Dec. 1976.

[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, August 1979.

[8] B. Yegnanarayana and Roel Smits, "A robust method for determining instants of major excitations in voiced speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 1995, vol. 1, pp. 776–779.

[9] Roel Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay functions," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 325–333, Sept. 1995.

[10] B. Yegnanarayana and N. J. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 4, pp. 313–327, July 1998.

[11] S. R. Mahadeva Prasanna, Suryakanth V. Gangashetty, and B. Yegnanarayana, "Significance of vowel onset point for speech analysis," in *Signal Processing and Communications (Biennial Conf., IISc Bangalore, India)*, July 2001, pp. 81–88.