SINGLE AND MULTICHANNEL ENHANCEMENT OF DISTANT SPEECH USING CHARACTERISTICS OF SPEECH PRODUCTION

B. Yegnanarayana[1], S. Guruprasad[2], S. R. Mahadeva Prasanna[3], and Suryakanth V Gangashetty[1]

[1]Speech and Vision Laboratory
International Institute of Information Technology (IIIT) Hyderabad - 500 032, India
Email: yegna@iiit.ac.in, svg@iiit.ac
[2]Department of Computer Science & Engineering
Indian Institute of Technology (IIT) Madras, Chennai - 600 036, India
Email: guru@research.iiit.ac.in
[3]Indian Institute of Technology (IIT) Guwahati - 781 039, India
Email: prasanna@iitg.ernet.in

ABSTRACT

Speech collected at a distance from a speaker is degraded due to background noise, reverberation and other audio/speech signals. Normally enhancement of the degraded signal focuses on determining the characteristics of degradation and auditory perception. There are many characteristics of speech production which contribute to robustness against degradation caused by distance between the speaker and the pickup microphone. One such characteristic is the sequence of impulse-like excitation of the vocal tract system during production of speech. This paper proposes single channel (mic) and multichannel (mics) enhancement methods for processing the distant speech. The single channel enhancement method exploits the pitch periodicity and the epoch locations to reduce the effect of additive background noise. A multichannel enhancement method is proposed not only to reduce the effect of background noise, but also to reduce the effects due to reverberation and other weak interfering signals. The multichannel method exploits the invariance property of the sequence of impulses due to excitation in the direct component of the sound at the two microphone locations. The results of enhancement are available at *http* : *//speech.iiit.ac.in/index.php/downloads.html/hscma*2011/.

Index Terms— Distant speech,reverberation, single channel enhancement, multichannel enhancement, epoch locations.

1. INTRODUCTION

Speech is perceived effortlessly even when there is a distance of over 1 m or more between speaker and listener. This is primarily due to the significance of the direct component of the speech at the receiver, and due to the binaural hearing ability and selective attention capability of the human listeners. When the signal is picked up by a microphone at a distance, and that signal is amplified and played back, it sounds degraded. The degradation in the speech signal is caused by the additive background noise, reverberation and other speech-like signals. The objective of this paper is to process the degraded distant speech signal to enhance the speech content.

Traditionally speech enhancement is viewed as reduction or suppression of the degrading component by suitably filtering the signal or subtracting the noise or dereverberating the signal. These approaches involve estimation of the spectrum of the degrading noise or estimation of the effect of reverberation [1, 2]. In implementing these enhancement methods, the perceptual effects of auditory processing are also taken into accounts. There are not many methods where the production characteristics of speech are exploited for enhancement of distant speech. It appears that speech is perceived even at a distance mainly because of the predominantly impulse-like excitation of the vocal tract system during production. The sequences of impulses are convolved with the response of the time-varying vocal tract system. The signal around the instant of impulse-like excitation corresponds to high signal-to-noise (SNR) region. Even at a distance the direct component of the speech signal from the speaker preserves the relative spacings and amplitudes of the excitation impulses. Also the signal around these impulses has higher SNR relative to other regions. Therefore the high SNR regions and their relative spacings are less affected by the additive background noise in the signal captured at a distance. On the other hand, the reverberant components of speech signal has lower amplitudes compared to the direct component. Also, due to several reflections, the relative spacings of the impulses are not preserved in the reverberant component. One can exploit the characteristics of the impulse-like excitation in the production of speech for enhancing the speech signal collected at a distance.

There are three types of degradation in distant speech: Additive background noise, reverberation and signals from background sources/speakers. In this paper we show that single channel signal can be enhanced if the degradation is predominantly due to additive noise. But for reverberation and interference due to other speech-like signals, two or more channel signals picked by spatially separated microphones are needed. The key idea is that the relative spacings of the excitation impulse sequences are preserved in the direct components at each of the microphones. By compensating for the fixed delay between a pair of microphone signals, the impulse sequences can be reinforced due to coherence. In is interesting to note that this not only reduces the degradation due to reverberation and other speech-like sources, but it also helps in significantly enhancing the speech against background noise, better than in the single channel case.

The paper is organized as follows: Section 2 describes the pro-

cedure for enhancing distance speech using single channel (microphone) data. Section 3 discusses the proposed approach for enhancement of distant speech using multichannel (microphones) data. Section 4 gives a summary of the paper and issues for further study.

2. SINGLE CHANNEL ENHANCEMENT OF DISTANT SPEECH

Speech signal collected at a distance can be categorized into three regions: (a) Region where the direct component of the signal is high relative to the reverberant component, (b) region where the reverberant component is dominant over the direct component, and (c) region containing mainly the reverberant component. With single channel data, the focus should be on emphasizing the direct component of the signal in regions where it is possible to do so, and on deemphasizing the reverberation only regions. The key idea of enhancement of distant speech is to emphasize the regions of high SNR in the excitation component of the speech signal. This involves the following steps.

(a) Voiced regions of speech are identified using the characteristics of zero-frequency filtered signal [3, 4].

(b) Within voiced regions, short segments around the instants of significant excitation are emphasized in the excitation component of the speech signal.

(c) The modified excitation signal is used to excite the all-pole filter represented by the linear prediction coefficients (LPCs).

2.1. Detection of voiced regions in distant speech

Characteristics of the impulse-like excitation of speech are exploited to detect the voiced regions in speech signals. Energy of the impulse-like excitation is concentrated in short intervals in time, and is uniformly distributed in frequency. The zero-frequency filtered signal displays significant amplitudes in the regions of glottal activity, relative to the regions of nonvoiced speech and silence [4]. The short-time energy of the filtered signal can be used for robust detection of voiced regions in speech signals. The short-time energy of the filtered signal is computed using segments of 30 ms each.

Figure 1(a) shows a segment of close-speaking speech signal. The corresponding distant speech signal, collected at 2 m from the speaker, is shown in Figure 1(d). It is interesting to note that the filtered signal (Figure 1(e)) derived from the distant speech signal has significant amplitudes only in those voiced regions where the direct component is stronger than the reverberant component. In the regions where only the reverberant component of speech is present (such as the tails of voiced sounds and the regions of silence), the filtered signal does not show any significant amplitude. For instance, the time interval between 300 ms and 350 ms in Figure 1 represents a region of silence. Yet, the amplitude of the distant speech signal (Figure 1(d)) in this region is comparable to that in the speech regions. By contrast, the filtered signal (Figure 1(e)) derived from the distant speech signal has relatively low amplitude in this region, compared to that in the speech regions. In the case of the distant speech signal, the voiced regions can be detected from the short-time energy of the filtered signal (Figure 1(f)). In the reverberation-only regions of the distant speech signal, the low amplitude values of the filtered signal are due to absence of strong impulse-like excitations in the reverberant components. The high SNR characteristic in short segments of the signal is absent in the reverberation-only regions.

Amplitude of the filtered signal is relatively low in the regions of speech signal where the glottal activity is absent. Such regions include speech sounds such as fricatives and unvoiced stop consonants. The use of short-time energy of the filtered signal for emphasizing the regions of voiced speech results in the deemphasis of these speech sounds. Information of such speech sounds may be partly preserved in the adjacent voiced segments due to effects of coarticulation. In addition to the coarticulation effects, the constraints imposed by the syntax and semantics of the spoken language also help in the perception of distant speech, even if some sounds are deemphasized. Nonlinear transformation is used to enhance the contrast between the values of the short-time energy of the filtered signal in the voiced regions, relative to the values in the nonvoiced regions, to derive a gross weight function [5].

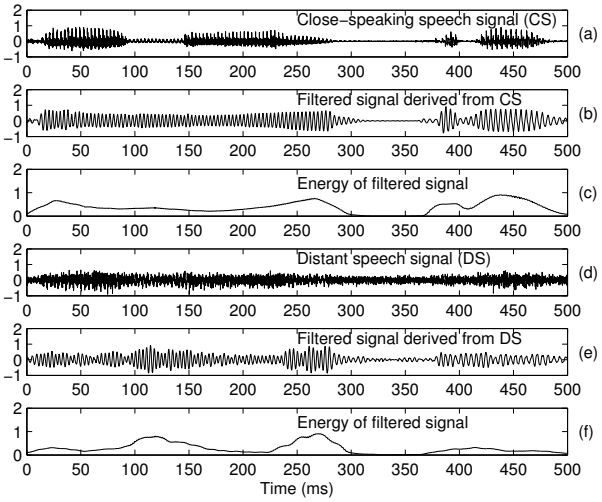

Fig. 1. Detection of voiced regions in distant speech signals: (a) A segment of close-speaking speech signal. (b) The corresponding filtered signal. (c) Energy of the filtered signal. (d) The segment of distant-speaking speech signal corresponding to the segment in (a). (e) Filtered signal derived from the distant speech signal. (f) Energy of the filtered signal.

A threshold can be applied on the values of the gross weight function so that the voiced regions in distant speech signal are identified. A small value of the threshold can result in a few nonvoiced and nonspeech regions being labeled as voiced regions. A large threshold can result in a few voiced regions being labeled as unvoiced regions. The choice of the threshold depends on the operation that is performed subsequently on the speech signal, based on the voiced/nonvoiced decision. The objective of the present study is to process the voiced regions which have been identified in the distant speech signal. Here, a smaller value of the threshold is preferred, so that voiced regions are not missed, even if a few nonvoiced regions are included.

2.2. Emphasis of regions around significant excitations in voiced regions

The pitch period information is derived from the distant signal using the algorithm proposed in [6]. The algorithm uses short-time Fourier transform of the filtered signal to detect the fundamental frequency (f_0). When the filtered signal is passed through an ideal digital resonator located at f_0, the resulting signal can help in detecting the epoch locations in the distant speech signal. From these epoch locations, the locations of the nearest peaks in the Hilbert Envelope

(HE) of the Linear Prediction (LP) residual are detected and are marked (by 'x') as shown in Figure 2.

Once these instants of significant excitation are detected, a time-varying function is constructed such that the function emphasizes short segments (2-3 ms) around the significant excitations, and deemphasizes the other regions between the significant excitations. This function is referred to as *fine weight function*, and is denoted by $w_f[n]$. An example of the fine weight function is shown in Figure 2(d). The fine weight function is used to emphasize the regions around the significant excitations in the LP residual, to obtain a modified LP residual. The advantage of this approach is that spurious impulse-like excitations present in the LP residual due to the effect of reverberation and noise are deemphasized. Deemphasis of spurious excitations can help in improving the perceptual quality of the synthesized speech. Since the fine weight function is constructed by using the instants of significant excitation as anchor points, the shape of the function can be controlled. The emphasis given to the regions around the significant excitations relative to the other regions can also be controlled. In the nonvoiced regions of the signal, the fine weight function $w_f[n]$ is set to a small value.

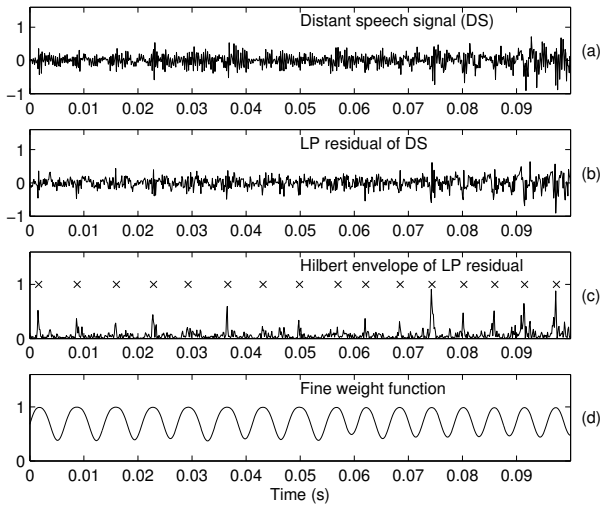

Fig. 2. Emphasis of regions around significant excitations. (a) A segment of distant speech signal. (b) The corresponding LP residual. (c) Hilbert envelope of the LP residual. Instants of significant excitations detected from the Hilbert envelope are marked by 'x'. (d) Fine weight function for emphasis of the regions around significant excitations in the LP residual.

Figure 3 shows the results of some of the steps involved in the derivation of the fine weight function. The gross weight function $w_g[n]$ derived from the short-time energy of the filtered signal is shown in Figure 3(b). Application of a threshold on the values of the gross weight function results in a voiced/nonvoiced decision shown in Figure 3(c). The fine weight function is shown in Figure 3(d). The modified LP residual $\tilde{e}[n]$ is given by $\tilde{e}[n] = w_f[n]e[n]$. The enhanced speech signal $\tilde{s}[n]$ is synthesized by exciting the all-pole filter represented by the set of LP coefficients with the modified LP residual $\tilde{e}[n]$. The LP coefficients obtained during the analysis stage remain unchanged.

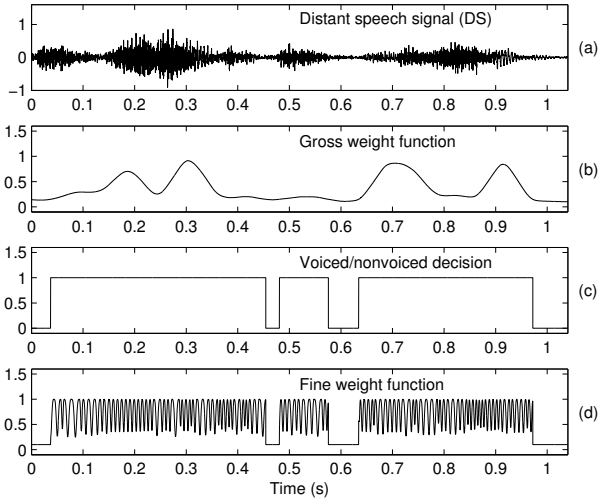

Fig. 3. Illustration of weight functions derived for emphasis of significant excitations. (a) A segment of distant speech signal. (b) Gross weight function derived from zero frequency filtered signal. (c) Voicing decision derived from the gross weight function. (d) Fine weight function used for emphasis of regions around significant excitations in the LP residual.

2.3. Results on single channel enhancement of distant speech

Speech signals were collected in a live room with dimensions of 8.5 m × 5.5 m × 3.0 m. The signals were collected at a distance of about 2 m from the speaker using a single-channel microphone. The environment included sources of noise such as air-conditioners and computer systems. The signals were sampled at 8 kHz. The collected speech signals were processed using the following three methods:

1. Source-based processing (SRC): The proposed method based on processing the excitation source signal is described in Sections 2.1 and 2.2.
2. Spectrum-based processing (SP): A method based on spectral subtraction is used [7].
3. Source-spectrum processing (SRC-SP): Speech signal is first processed using the source-based method. The output signal is then processed using the spectrum-based method.

The objective of source-spectrum processing is to exploit the advantages of the source-based and the spectrum-based methods. In the source-based method of processing, detection of voiced regions helps in reducing the noise power in nonvoiced regions. This reduces the criticality of estimation of noise power in the spectrum-based method of processing.

Figure 4 shows a distant speech signal and the corresponding processed signals. Figure 4(c) shows the signal obtained by the source-based method of processing. The corresponding spectrogram is shown in Figure 4(d). Reduction of energy in the nonvoiced segments relative to that in the voiced segments can be observed in both time and frequency domains. Since the source-based method does not process the short-time spectrum of the signal, formants in voiced regions of the processed signal are not emphasized relative to the spectral valleys (Figure 4(d)). In the signal obtained by the spectrum-based method of processing (Figure 4(e)), reduction of energy in the noisy segments is more pronounced, as observed from the spectrogram (Figure 4(f)). Duration of the signal used for estimating

the spectral characteristics of noise is an important parameter. The duration must be long enough to obtain an accurate estimate of the spectral characteristics of noise, but short enough to track (nonstationary) variations of noise. In this case, the spectral characteristics of noise are estimated using signal windows of size 1.5 s, and spectral subtraction is performed on subsequent segments of speech. Hence, the initial segment of 1.5 s in the processed signal is noisy (Figure 4(f)). The signal obtained using source-spectrum processing is shown in Figure 4(g). In this case, spectral characteristics of noise are estimated from the signal obtained by source-based processing (Figure 4(c)). Reduction of energy in the noisy segments in Figure 4(h) is better than that in Figure 4(f).

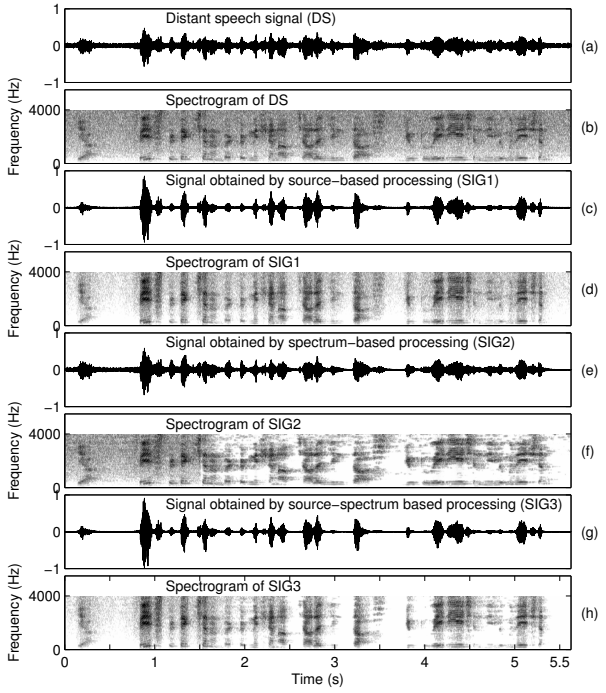

Fig. 4. Illustration of enhancement of distant speech signal. (a) A segment of distant speech signal and (b) the corresponding spectrogram. (c) Signal obtained by source-based processing and (d) the corresponding spectrogram. (e) Signal obtained by spectrum-based processing and (f) the corresponding spectrogram. (g) Signal obtained by source-spectrum based processing and (h) the corresponding spectrogram.

Subjective studies were conducted to assess the performance of the three methods of processing. Two English sentences uttered by a male speaker were used for evaluation. For each sentence, the four signals (the original speech signal and the three processed signals) were normalized such that they had the same overall energy. Ten subjects participated in the listening test to evaluate the three methods of enhancement. The subjects rated the processed speech signals on the basis of two measures, namely, *distortion* and *overall perceptual quality*. The subjects rated the distortion on a four-point scale ranging from 1 to 4. A rating of 1 indicates no distortion, while a rating of 4 indicates significant distortion. Similarly, the overall perceptual quality was rated using a five-point scale ranging from 1 to 5. A rating of 1 indicates that the processed signal is perceived to be worse than the original signal, while a rating of 5 indicates that the processed signal is perceived to be significantly better than the original signal. Distortion was judged in terms of factors such as perceptual discontinuities, unnaturalness and synthetic quality. The overall perceptual quality was judged on the basis of factors such as reduction in noise/degradation, improvement in intelligibility, comfort of listening and tolerability of distortion. For each sentence, the original (collected) speech signal served as the reference, on the basis of which the three processed signals were judged. For each sentence, the identity of the original (collected) speech signal was known to the subjects, but no information about the processing methods was provided to them. For the three methods, namely SRC, SP and SRC-SP, the mean distortion scores were observed to be 1.83, 2.72 and 2.33, respectively. The spectrum-based method introduces musical noise, due to subtraction of average noise spectrum from each frame of the signal. Hence, the processed signal is perceived to be more distorted than that obtained from source-based method. The mean perceptual quality scores for SRC, SP and SRC-SP methods were observed to be 2.94, 3.75 and 4.16, respectively. Since the spectrum-based method has a 'cleaning-up' effect on the signal, the output is perceived to be less noisy than the original signal. In the source-spectrum based method, the source processing reduces the amplitude of nonspeech regions, which in turn reduces the criticality of estimation of noise spectrum in the subsequent spectral subtraction. Hence the mean perceptual quality score of SRC-SP method is higher than that of SP method.

3. MULTICHANNEL ENHANCEMENT OF DISTANT SPEECH

When speech is transmitted in an acoustical environment like in an office room, it will be degraded by background noise and reverberation [5, 8–12]. Multichannel case is more effective for enhancement compared to the single channel case, but requires estimation of time-delays [8]. One simple method for enhancement in multichannel case is addition of the speech signals, after compensating for their delays. Coherent addition of speech signals from different microphones will provide enhancement mainly against background noise. The improvement in enhancement is directly related to the number of microphones used. For achieving significant enhancement, especially due to reverberation, additional processing of the microphone signals is required. In this section a method for enhancement using the epoch information is described.

3.1. Coherent Addition of Speech Signals

Speech was collected from 14 spatially distributed microphones placed in an office room of dimension 3 m × 4 m × 3 m with a reverberation time of about 200 ms. The delay between every pair of microphones is computed using the excitation source information. The coherently-added signal, obtained after compensating for their delays is given by [13]

$$s_{e1}(n) = \frac{1}{N}[s_1(n) + s_2(n - \tau_{12}) + ... + s_N(n - \tau_{1N})] \qquad (1)$$

where τ_{1i} is the delay in samples between mic-1 and mic-*i*.

The coherent addition reinforces speech components and thus reduces the effect of the background noise. However, the reverberant component is still present in the resulting signal. The degree of enhancement achieved at this level depends on the number of microphones used in the coherent addition. Figures 5(a) and (b) show the clean speech signal, and the degraded speech signal collected from *mic-1*, respectively. Figures 5(c) and (d) show the enhanced speech

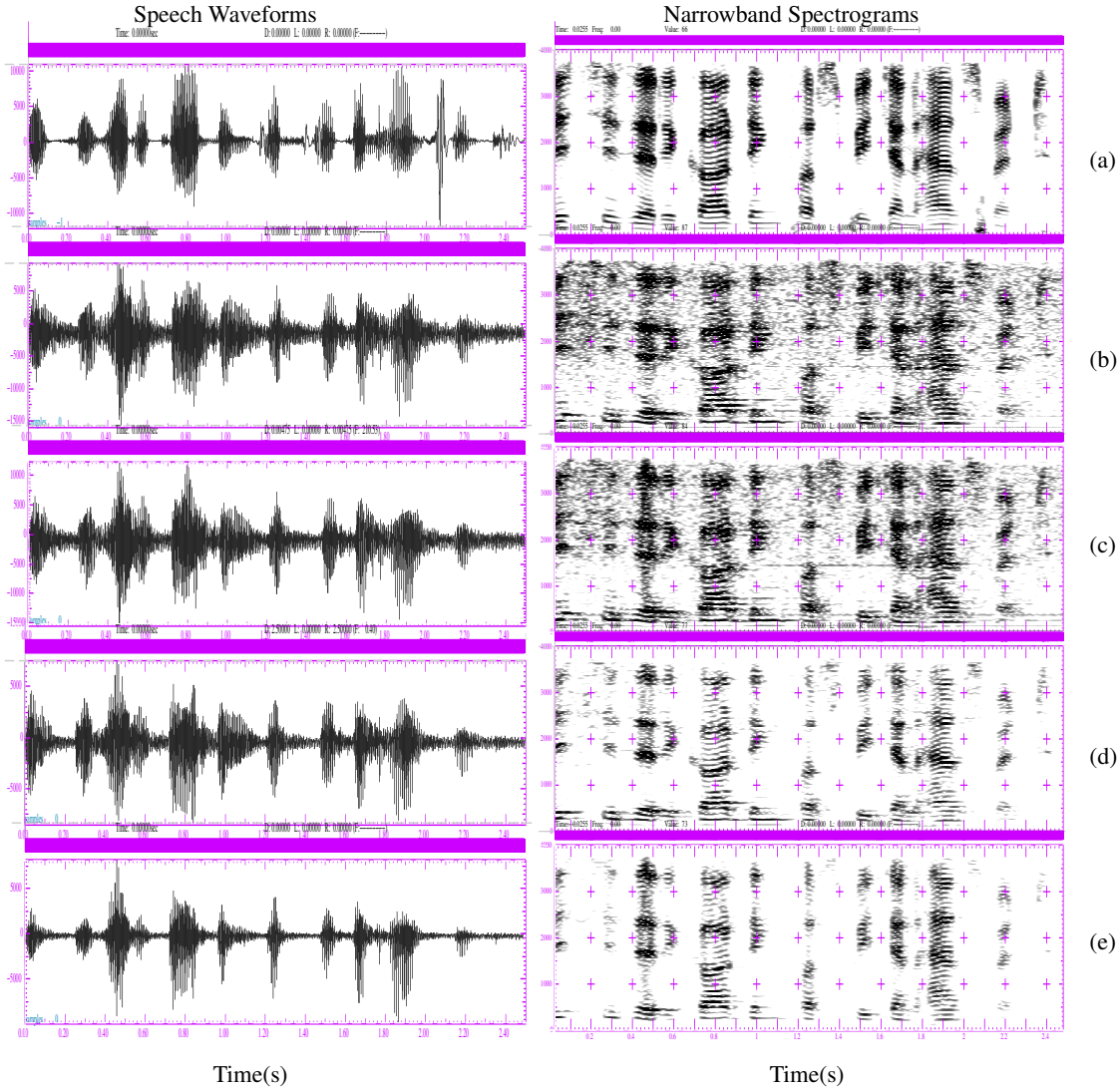

Fig. 5. Each row shows a speech signal and its narrowband spectrogram for: (a) Close speaking microphone. (b) Distant microphone (*mic-1*). (c) and (d) Enhanced signals obtained by coherent addition of two microphone signals and fourteen microphone signals, respectively. (e) Enhanced signal using the coherently-added Hilbert envelope of the LP residuals of three microphone signals [13].

signals and their narrowband spectrograms, obtained by the coherent addition of signals of two microphones and fourteen microphones, respectively. As can be seen from the narrowband spectrograms, there is decrease in the background noise as we increase the number of microphones. It is interesting to note that in the case where signals from fourteen microphones are coherently added, the reverberation tails are present in the speech regions, even though the effect of background noise is reduced. This can be observed by comparing the spectrogram, especially at low frequencies, with that of the clean speech shown in Figure 5(a). The presence of reverberation tails is also clearly visible in the waveforms (compare Figure 5(a) and Figure 5(d)). It is necessary to process the coherently-added speech signal further to achieve enhancement with respect to reverberation.

3.2. Processing of Excitation Source Information

For each of the microphone signals, the Hilbert envelope of the LP residual is computed. The coherent addition of the Hilbert envelopes reinforces the epoch information, whereas the incoherent addition will spread the epoch information. For coherent addition, the delay between two microphone signals is computed using the cross-correlation of the HEs of the LP residuals of the microphone signals [13]. The coherently-added Hilbert envelope exhibits several interesting features. The deviation among the samples of the Hilbert envelopes is high in the voiced speech regions. Typically, voiced speech regions in continuous speech have a minimum duration of 50 ms. Hence, by considering a block of 50 ms duration and a shift of one sample, the mean and standard deviation of the coherently-added Hilbert envelope samples in each block are computed. The standard deviation values are normalized with the respective mean values. In the normalized standard deviation plot, the deviation of

the Hilbert envelope samples is high in the speech regions (see Figure 6(a)). Further, the normalized standard deviation is high in the initial portions of the voiced speech regions, and it decreases towards the end of the voiced regions. This is because the initial parts are high SNR regions. Towards the end of the voiced regions, the levels of degrading components increase, and hence they correspond to low SNR regions. Another interesting property of the coherently-added Hilbert envelope is that, the samples in each pitch period around the epochs have large deviation compared to the samples away from the epoch, when the mean, standard deviation and normalized standard deviation are computed for a block size of 3 ms and a shift of one sample (see Figure 6(b)).

A weight function is derived by adding the two (long and short blocks) normalized standard deviation values as shown in Figure 6(c). The LP residual of one microphone is weighted using the weight function. The weighted residual is used to excite the time varying all-pole filter derived from the coherently-added signal, to synthesize the enhanced speech signal. Figure 5(e) shows the enhanced speech signal obtained by processing the LP residual of distant speech. In this case, Hilbert envelope signals of three channels were coherently added, to derive the weight function. From Figure 5(e), it can be seen that the speech signal is enhanced both with respect to background noise as well as with respect to reverberation.

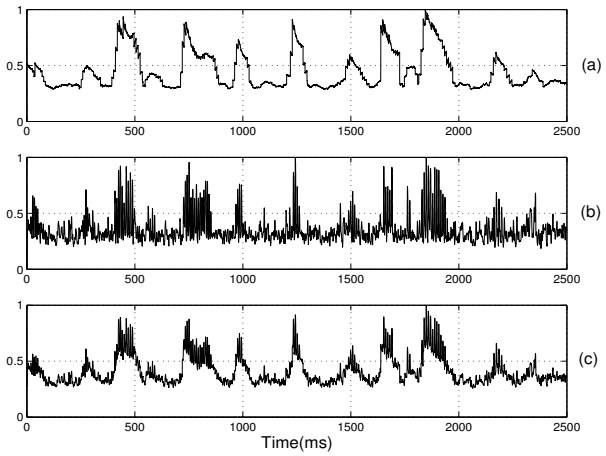

Fig. 6. (a) Normalized standard deviation plot derived using block size of 50 ms and shift of 1 sample, (b) normalized standard deviation plot derived using block size of 3 ms and shift of 1 sample and (c) weight function obtained by adding (a) and (b) [13].

4. SUMMARY AND CONCLUSIONS

This paper proposed methods for enhancing the distant speech by exploiting the characteristics of the excitation source in speech production. The impulse-like excitation in the voiced regions of speech provide robustness for perception of speech even at a distance. The regions around these impulse-like excitations generally have higher SNR, and the spacing between impulses is preserved in the direct component of the speech at the microphones. The high SNR regions are emphasized relative to the other regions to enhance distance speech. The region of glottal activity and the epoch locations are determined using zero-frequency filtered signal for the single channel enhancement case. In multichannel case, it is possible to achieve higher noise suppression, besides reducing significantly the effects of reverberation and other interfering sounds. This is achieved by the coherent addition of the impulse excitation regions in the delay compensated Hilbert envelope of the LP residuals of the microphone signals. Note that the enhancement is achieved by mainly focusing on the excitation component. The system characteristics (LPCs) derived from the degraded signals are used as they are. Hence there is scope for the improvements in the enhancement by modifying the system characteristics also.

5. REFERENCES

[1] Michael Kleinschmidt, Jurgen Tchorz, and Birger Kollmeier, "Combining speech enhancement and auditory feature extraction for robust speech recognition," *Speech Communication*, vol. 34, no. 1-2, pp. 75–91, Apr. 2001.

[2] A. Sehr, R. Maas, and W. Kellermann, "Reverberation model-based decoding in the logmelspec domain for robust distant-talking speech recognition," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 18, no. 7, pp. 1676–1691, Sep. 2010.

[3] K. Sri Rama Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.

[4] K. S. R. Murty., B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal Process. Letters*, vol. 16, no. 6, pp. 469–472, June 2009.

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

[6] Guruprasad S. and B. Yegnanarayana, "Performance of an event-based instantaneous fundamental frequency estimator for distant speech signals," *IEEE Trans. Audio, Speech Lang. Process.*, accepted for publication, 2010. DOI: 10.1109/TASL.2010.2101595.

[7] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. European Signal. Process. Conf.*, Edinburgh, UK, Sep. 1994, pp. 1182–1185.

[8] J. L. Flanagan, J. D. Jonston, R. Zahn, and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *Journal of Acoustical Society of America*, vol. 78, no. 5, pp. 1508–1518, 1985.

[9] B. Yegnanarayana, C. Avendaño, H. Hermansky, and P. S. Murthy, "Speech enhancement using linear prediction residual," *Speech Communication*, vol. 28, no. 1, pp. 25–42, May 1999.

[10] J. Jensen and J. H. L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 7, pp. 731–740, Oct. 2001.

[11] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[12] J. Huang and Y. Zhao, "An energy-constrained signal subspace method for speech enhancement and recognition in colored noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Seattle, WA, USA, May 1998, pp. 377–380.

[13] S. R. Mahadeva Prasanna, "Event Based Analysis of Speech," PhD Thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Mar. 2004.