Enhancement of Speech in Multispeaker Environment

B. Yegnanarayana, S.R. Mahadeva Prasanna

Speech and Vision Laboratory
Department of Computer Science and Engg.,
IIT Madras, Chennai-600 036, India
Email:{yegna,prasanna}@cs.iitm.ernet.in

Mathew Magimai Doss

Dalle Molle Institute for Perceptual Artificial Intelligence
(IDIAP), CH-1920, Martigny, Switzerland, and
The Swiss Federal Institute of Technology (EPFL),
CH-1015, Lausanne,Switzerland
Email:mathew@idiap.ch

Abstract

In this paper a method based on the excitation source information is proposed for enhancement of speech, degraded by speech from other speakers. Speech from multiple speakers is simultaneously collected over two spatially distributed microphones. Time-delay of each speaker with respect to the two microphones is estimated using the excitation source information. A weight function is derived for each speaker using the knowledge of the time-delay and the excitation source information. Linear prediction (LP) residuals of the microphone signals are processed separately using the weight functions. Speech signals are synthesized from the modified residuals. One speech signal per speaker is derived from each microphone signal. The synthesized speech signals of each speaker are combined to produce enhanced speech. Significant enhancement of the speech of one speaker relative to other was observed from the combined signal.

1. Introduction

In a multispeaker environment, speech signal may be corrupted by the presence of additive noise, reverberation, speech of other speakers or a combination of all of these. The quality of the degraded speech will be poor for listening, and also the performance will be poor when features are extracted for various applications like source localization, tracking a moving speaker, speech recognition, speaker recognition, and audio indexing. Therefore there is a need for enhancing the desired speaker's voice from the degraded speech before using it for any application. This paper proposes a method for enhancing speech degraded by speech from other speakers. The objective is to produce perceptually enhanced speech by processing the degraded speech.

Several methods have been proposed in the literature for enhancement of speech corrupted by speech from other speakers [1–6]. These methods may be broadly classified into two categories, namely, single channel case and multichannel case. In the single channel case, speech is collected over a single microphone, and the objective is to process the speech signal to emphasize the desired speaker's voice. This approach is more commonly termed as cochannel speaker separation [2, 3]. The implicit assumption in most of the proposed single channel methods is that there are only two speakers, and among them one is the desired speaker. In the multichannel case, speech is collected simultaneously over several (two or more) spatially distributed microphones. Signals from all the microphones are processed to enhance the speech of one or more of the speakers. This approach seems to be inspired by the binaural method of speech processing present in human beings [4]. The proposed method uses speech collected simultaneously over two spatially distributed microphones (say *mic-1* and *mic-2*), and the speech is processed for enhancement.

Most of the existing methods for enhancement are based on the vocal tract system features [3, 5]. This paper proposes a method based on the excitation source information to produce perceptually enhanced speech [7]. The glottal closure (GC) event or epoch and a small region around the GC event corresponds to the high signal-to-noise ratio (SNR) region in each pitch period of the speech signal [7, 8]. Hence identifying and enhancing such regions for each speaker will produce perceptually enhanced speech. As the speakers are at different distances with respect to the spatially distributed microphones, there will be an unique sequence of epochs associated with each speaker in the microphone signals. This sequence may be enhanced in the Linear prediction (LP) residuals of the microphone signals using a suitable weight function. The modified residual is used to excite the time-varying all-pole filter to synthesize the enhanced speech.

The paper is organized as follows: Section 2 discusses the excitation source information in the speech signal, and the basis for the proposed method of enhancement. In Section 3 the method for enhancement based on excitation source information is described. Experimental results are given in Section 4. The paper is concluded in Section 5 with a summary of the present work and possible extensions for further improvement.

2. Excitation Source Information for Speech Enhancement

Speech is the result of excitation of a time-varying vocal tract system with time-varying excitation. The shape of the vocal tract decides the type of the sound unit to be produced, and the strength of excitation provides the required energy. It is interesting to note that the GC event and the high SNR region around the GC event provide most of the energy for the speech signal. This can be observed by replacing the glottal excitation by random noise, which results in the production of whispered speech, that can hardly be perceived at a distance.

The vocal tract system information is reflected in the spectral features. Hence after removing the spectral features from the speech signal, the residual contains mostly information about the excitation source. This separation can be achieved by LP analysis [9]. In LP analysis each sample is predicted as a linear combination of the past p samples, where p is the order of prediction. The predicted sample is given by

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \qquad (1)$$

where $\{a_k\}$ are the LP coefficients (LPCs). The LPCs are obtained by solving the set of p normal equations

$$\sum_{k=1}^{p} a_k R(n-k) = -R(n) \qquad n = 1, ...p \qquad (2)$$

where

$$R(k) = \sum_{n=0}^{N-(p-1)} s(n)s(n-k) \qquad k = 0, ...p \qquad (3)$$

and $\{s(n)\}$ are the speech samples. The filter designed using $\{a_k\}$ models the vocal tract system information. The inverse filter is given by

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \qquad (4)$$

The LP residual which mostly contains the excitation source information is derived by passing the speech signal through the inverse filter.

The LP residual may be processed for accurately detecting the GC events as proposed in [10]. In the present task, the objective is only to identify the high SNR region which is around the GC event. Hence accurate location of the GC event is not required and it is enough to locate the region around the GC event. This can be achieved using the Hilbert envelope of the LP residual, which is defined as

$$h(n) = \sqrt{r^2(n) + r_h^2(n)} \qquad (5)$$

where $r_h(n)$ is the Hilbert transform of the LP residual $r(n)$, and is given by

$$r_h(n) = IDFT[jDFT[r(n)]] \qquad (6)$$

The DFT and $IDFT$ are the discrete and inverse discrete Fourier transforms, respectively [11]. Figure1 shows a segment of voiced speech, the LP residual, the GC events, and the Hilbert envelope of the LP residual. The peaks in the Hilbert envelope give the region around the GC events.

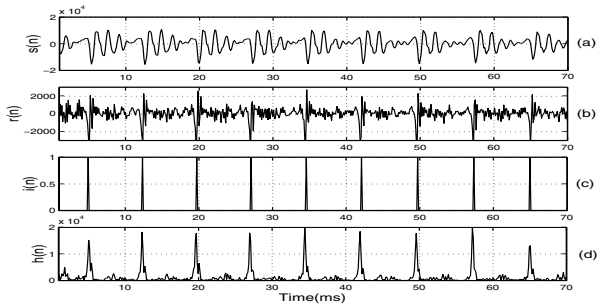

Figure 1: (a) Voiced segment of speech, and its (b) LP residual, (c) GC events and (d) Hilbert envelope.

The first step in processing multichannel speech data is the estimation of the time-delay. Time-delay is computed using the cross-correlation of the Hilbert envelopes of the LP residuals of two microphone signals. For instance, typical computed delays for every frame of 100 ms with a shift of 5 ms are shown in Figure2. As the number of speakers in the present study are two, there are mainly two delay values (say, *TD1* and *TD2*), one for each speaker (say, *SP1* and *SP2*). Most of the random delays correspond to nonspeech regions.

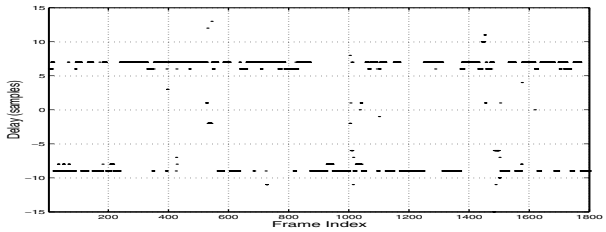

Figure 2: Time-delays computed from the Hilbert envelope of two microphone signals for two speaker case.

Segments of the Hilbert envelopes of the LP residuals of individual microphones are shown in Figures3(a) and 3(b). These are added after compensating for each of the two delays, and are shown in Figures3(c) and 3(d). For each delay, there are instants at which the samples of the Hilbert envelopes will be added coherently. These instants are the locations of the glottal excitations for the speaker corresponding to that delay. This property of reinforcement of the excitation information in the combined Hilbert envelope for each speaker is exploited to

derive a weight function for enhancing the high SNR portions of the excitation, mostly belonging to that speaker.

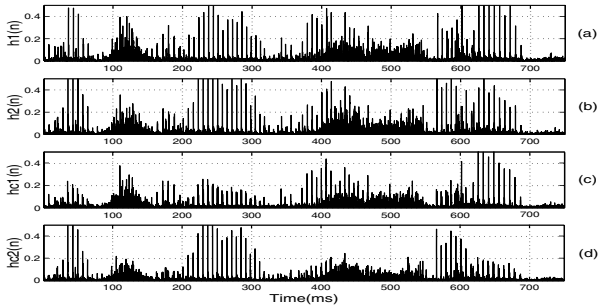

Figure 3: Hilbert envelopes of (a) *mic-1* LP residual and (b) *mic-2* LP residual. Combined Hilbert envelopes using (c) *TD1* and (d) *TD2*.

3. Proposed Method for Enhancement

To derive a weight function for each speaker, mean and standard deviation values are computed by considering a frame size of 2 ms and a frame shift of one sample for the Hilbert envelopes shown in Figure3. The standard deviation values are divided by the corresponding mean values and are shown in Figures4(a) to 4(d). The normalized standard deviation values in Figures4(c) and 4(d) are divided by the average of corresponding values (after shifting) in Figures4(a) and 4(b). These ratios are shown in Figures4(e) and 4(f), corresponding to Figures4(c) and 4(d), respectively. These ratios are close to one in the regions where the samples of the Hilbert envelopes are coherently added. These regions correspond to the high SNR regions of the speaker whose samples of the Hilbert envelope are coherently added. Weight functions can be derived from these ratios using a suitable threshold value, which is 0.8 in the present study, and are shown in Figures4(g) and 4(h). The weight functions are processed further to minimize the distortion due to their rectangular shape.

The LP residuals are weighted using each of the weight functions. The modified residuals are used to excite the time-varying all-pole filter obtained using the LPCs derived from the original microphone signals. One signal per speaker is synthesized from each microphone signal. The synthesized speech signals of a speaker are coherently added to produce the enhanced speech.

4. Experimental Results

Speech data for this study was taken from ICA99 speech signals for blind signal separation [12]. The speech signals were collected over two omnidirectional microphones in a room with dimension $3.1 \times 4.2 \times 5.5$ m, having some background noise during recording. The signals were sampled at 16 kHz and stored with 16 bit resolution. The

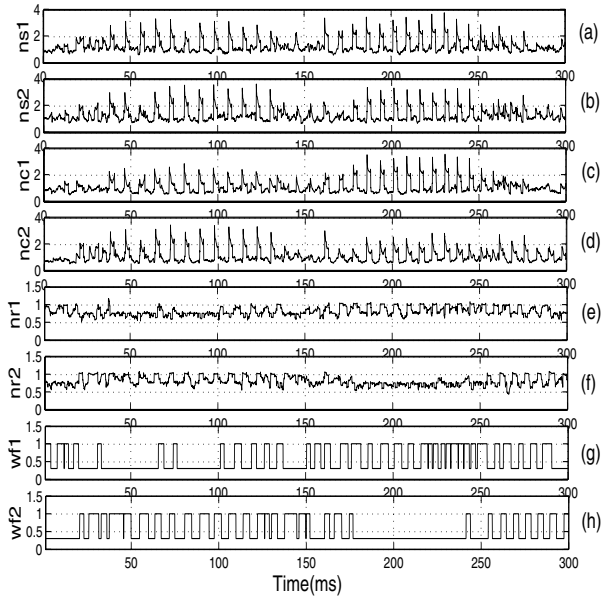

Figure 4: Normalized standard deviation plots of (a) *mic-1* Hilbert envelope and (b) *mic-2* Hilbert envelope. Combined Hilbert envelope using (c) *TD1* and (d) *TD2*. Ratio of normalized standard deviations for (e) *TD1* and (f) *TD2*. Weight function for (g) *TD1* and (h) *TD2*.

speech signals are preemphasized, and the LP residuals are extracted for every frame of 20 ms with a shift of 10 ms using an LP order of 20. Weight functions are derived as described in the previous section. The residuals are modified using the weight functions and the speech signals are synthesized from the modified residuals.

Figures5(a) and 5(b) show the speech signals of the individual microphones. Figures5(c) and 5(e) show the enhanced speech signals of *SP1* and *SP2*, respectively, as given in the database for comparison. The enhanced speech signals of *SP1* and *SP2* by the proposed approach are shown in Figures5(d) and 5(f), respectively. The enhanced speech signals from the proposed approach are better than the given reference signals, especially in the high voiced regions.

The wideband spectrograms are shown in Figure6 for each of the speech segments shown in Figure5. As can be seen from these figures, the proposed method enhances speech of each speaker, with no spectral distortion. Perceptually it was found that the separation of the speaker by the proposed approach is better compared to the reference signals. All the speech files used in this study can be downloaded from the site:
http://speech.cs.iitm.ernet.in/Results/MultiSpkr.html.

5. Conclusions

A method was proposed for enhancement of speech in multispeaker environment. The unique delay between

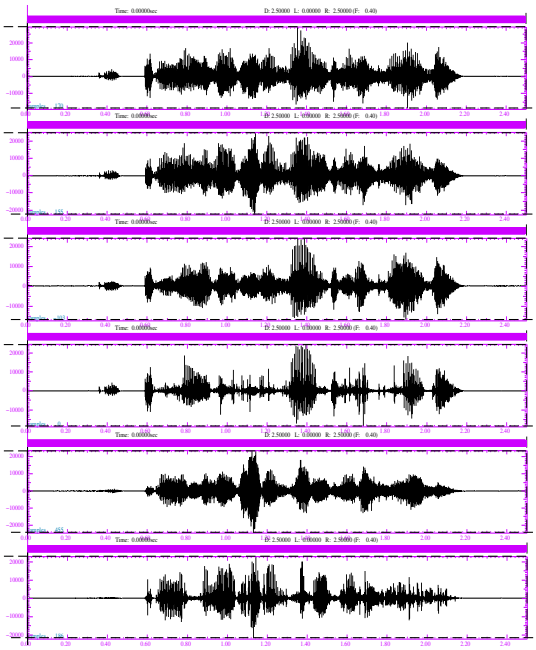

Figure 5: Speech signal of (a) *mic-1*, (b) *mic-2*, (c) *SP1* reference speech signal, (d) *SP1* speech signal by proposed method, (e) *SP2* reference speech signal and (f) *SP2* speech signal by proposed method.

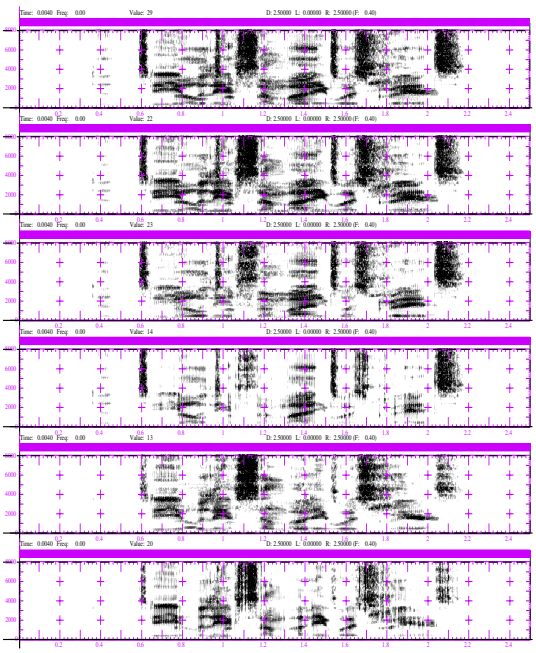

Figure 6: Spectrograms of the speech signal of (a) *mic-1*, (b) *mic-2*, (c) *SP1* reference speech signal, (d) *SP1* speech signal by proposed method, (e) *SP2* reference speech signal and (f) *SP2* speech signal by proposed method.

pair of microphones and the knowledge of excitation source information are exploited to develop a method to enhance the high SNR regions to produce perceptually enhanced speech. This study uses LP coefficients derived from the original microphone recordings, and thus contains information about both the speakers. It is necessary to obtain LPCs corresponding to each speaker to provide better enhancement in the spectral domain. Processing the spectral as well as the source features may provide a better method for enhancement of speech in a multispeaker environment.

6. References

[1] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Amer.*, vol. 60, pp. 911–918, 1976.

[2] C.K. Lee and D.G. Childers, "Cochannel speech separation," *J. Acoust. Soc. Amer.*, vol. 83(1), pp. 274–280, 1988.

[3] D.P. Morgan, E.B. George, L.T. Lee, and S.M. Kay, "Cochannel speech separation by harmonic enhancement and supression," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 407–424, 1997.

[4] O.M.M. Mitchell, C.A. Ross, and G.H. Yates, "Signal processing for a cocktail party effect," *J. Acoust. Soc. Amer.*, vol. 50, pp. 656–660, 1971.

[5] J-F. Cardoso, "Blind signal separation: statistical principles," *Proc. IEEE*, vol. 86, pp. 2009–2025, 1998.

[6] A.K. Barros, T. Rutkowski, F. Itakura, and N. Ohnishi, "Estimation of speech embedded in a reverberant and noisy environment by independent component analysis and wavlets," *IEEE Trans. Neural Networks*, vol. 13, pp. 889–893, 2002.

[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*, 2002.

[8] 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, 2000.

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

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

[11] 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, 1979.

[12] ICA'99, "Int. workshop on independent component analysis and blind signal separation," in *http://www2.ele.tue.nl/ica99/realworld.html*, 1999.