SEPARATION OF MULTISPEAKER SPEECH USING EXCITATION INFORMATION

B. Yegnanarayana[1], R. KumaraSwamy[1] and S. R. Mahadeva Prasanna[2]

[1]Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600 036 India
email: {yegna, kswamy}@cs.iitm.ernet.in
[2]Department of Electronics and Communication Engineering
Indian Institute of Technology Guwahati, Guwahati - 781 039 India
Email:prasanna@iitg.ernet.in

Abstract. In this paper, we propose an approach for separating speech of individual speakers from a multispeaker speech signal using excitation source information. The proposed approach is demonstrated in a two-microphone case. The main issue in the two-microphone case is the estimation of delay of each speaker. We propose a method for delay estimation in multispeaker case using the knowledge of excitation source information. The estimated delays are used for deriving weight functions for each speaker. The weight functions are used for extracting the excitation sequences for each of the speakers. The separated speech for each speaker is synthesized using the extracted excitation sequence. The proposed approach is illustrated for three speaker speech data collected over two spatially distributed microphones.

1 Introduction

One of the challenging problems in signal processing is separation of speech due to each individual speaker from a speech signal collected by a microphone when several speakers are speaking simultaneously [1–3]. The problem is compounded by the fact that the signal is corrupted by both additive (environmental) noise and room reverberation. The only clue one may have for this separation problem is the difference in the characteristics of each individual voice. But it is almost impossible to determine the characteristics of the individual voices from the combined speech collected by a microphone.

The problem becomes less complicated if the speech signals are collected simultaneously at two or more spatially distributed microphones [3, 4]. In such a case the feature one could exploit is the delay in the speech signals produced by an individual at any two microphone locations. The delays are different for each speaker, as no two speakers can be exactly at the same location. Some special and trivial cases such as locations along the perpendicular bisector of the line joining the two microphones are ignored for the time being, as such cases can easily be handled by using a large number of spatially distributed microphones.

In the present study the data recording scenario involves collection of speech produced by three or more speakers speaking simultaneously using two microphones separated by about 1.5 m. The microphones are approximately 1 to 2 m from the speakers. The recording is done in a laboratory environment with associated background noise and reverberation. The locations of speakers and microphones are fixed throughout the recording, so that all the delays are constant. The speech data is sampled at 8 kHz, and hence accuracy of the delay is limited to the sampling interval.

There are several methods proposed in the literature for multiple speaker separation. Most of them use either blind deconvolution or Independent Component Analysis (ICA) methods [3, 5–7]. Moreover, the methods rely on the spectral characteristics of the signal and the differences in the fundamental frequencies of the individual speakers. In this paper we propose a method based on the information in the excitation characteristics of speech production. Speech is produced as a result of exciting time varying vocal tract system with time varying excitation. The significant excitation is mostly due to impulse-like excitation caused at the onset of bursts and around the instants of glottal closure in voiced speech. Thus the excitation of the vocal tract system may be considered as a sequence of impulses, located at random instants for non-voiced sounds and at regular quasi periodic instants for voiced sounds. The relative positions of these impulses in the speech remain constant at any microphone location in the room. Only delay between two microphones will change depending on the locations of the microphones in the room.

The paper is organized as follows: In Section 2, we discuss the issues in the time delay estimation of individual speakers using the speech signals at the two microphones. The processing of excitation source information using time delays and deriving the weight function to extract the excitation information of individual speakers is discussed in Section 3. The results of speaker separation from a three speaker speech signal are also discussed in this section. Section 4 gives a summary of this work.

2 Time delay estimation

Estimation of the delay between two microphone signals can be done better in the excitation component of the speech signal, rather than in the speech signal itself [8–10]. For this estimation even an approximation to the excitation component may be useful. Hence in this paper we propose the use of residual derived from the speech signal by Linear Prediction (LP) analysis [11]. The LP residual is the error between the speech signal and its predicted value and is given as $e(n) = s(n) - \hat{s}(n)$, where $\hat{s}(n)$ is the predicted value of $s(n)$. We use a 12^th order LP analysis on each 20 msec frame of speech data, using a frame shift of 5 msec between successive frames. The LP residual has large error around the onset of bursts and around the instants of glottal closure. But the polarity of the samples of the residual signal vary at each of the these instants. Therefore we propose to use the Hilbert envelope of the residual signal, which shows strong

peaks around the excitation impulses. The Hilbert envelope is the magnitude of the analytic signal derived from the LP residual [12, 13]. The analytic signal of the LP residual is given as $e(n) + je_h(n)$, where $e(n)$ is the residual and $e_h(n)$ is the Hilbert transform of $e(n)$. The Hilbert envelope of the LP residual is defined as $h(n) = \sqrt{e^2(n) + e_h^2(n)}$. Fig. 1 shows the speech signal, its LP residual and the Hilbert envelope of the LP residual.

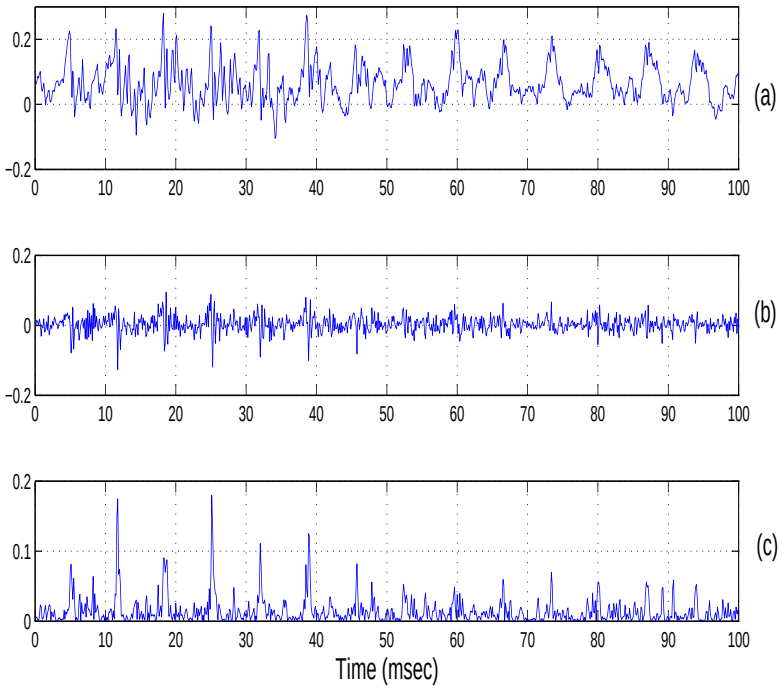

Fig. 1. (a) Speech signal, (b) LP residual, and (c) Hilbert envelope of the LP residual.

Fig. 2 shows the Hilbert envelopes of the LP residuals of the speech signals collected at two microphones (*mic-1* and *mic-2*). The speech signal is the speech produced by three speakers speaking simultaneously. The delay between the signals of each speaker is obtained by computing the cross correlation function of the Hilbert envelopes of the two microphone signals. Cross correlation is performed using a frame size of 100 msec and a frame shift of 5 msec [8, 9]. Fig. 3 shows the locations of peaks in the cross correlation as a function of the frame index. Each frame index corresponds to the shift of 5 msec. Fig. 4 shows the normalized plot of the number of points for each delay. The three strong peaks correspond to the time delays between the microphones due to each speaker.

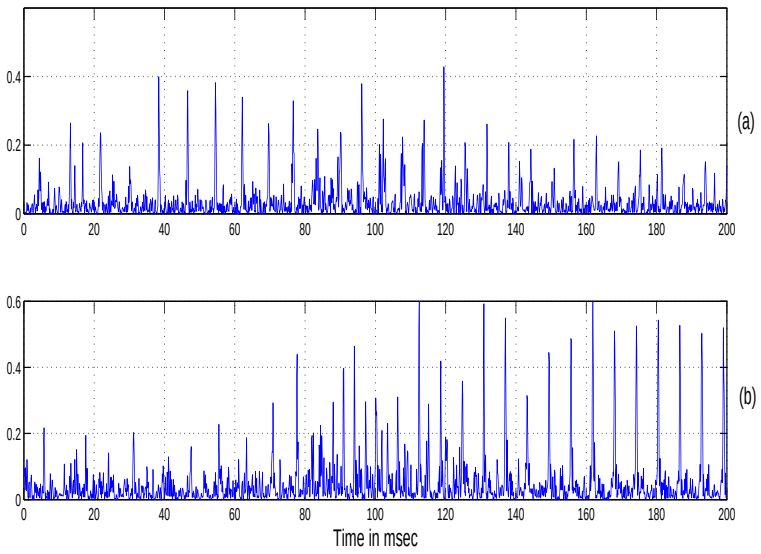

Fig. 2. Hilbert envelopes of the LP residual of (a) *mic-1* and (b) *mic-2* signals.

3 Processing excitation source information

Using the estimated delays and the Hilbert envelopes of the LP residuals, one can separate the excitation information corresponding to each speaker as follows. Keeping the Hilbert envelope of *mic-1* as reference, the Hilbert envelope of *mic-2* is shifted by one of the three delays, and the minimum of the two Hilbert envelopes is obtained. The minimum function obtained for each speaker using the respective time delays are obtained as shown in Fig. 5. The relatively high values in the minimum function indicate the significant excitation regions of the desired speaker.

The minimum function of each speaker is processed further to derive a weight function which gives more emphasis to the desired speaker. The weight function is used to multiply the LP residual of *mic-1* to obtain the modified residual for each speaker. The modified residual is used to excite the time varying all-pole filter derived from the *mic-1* signal. Note that the all-pole filter is derived from the combined signal. The weights of the residual emphasizes the desired speaker even though the spectral characteristics are not changed. Fig. 6 shows the LP residual of *mic-1* and the modified residuals for the 3 speakers. The separated speech signals by the proposed approach are also shown in Fig. 6.

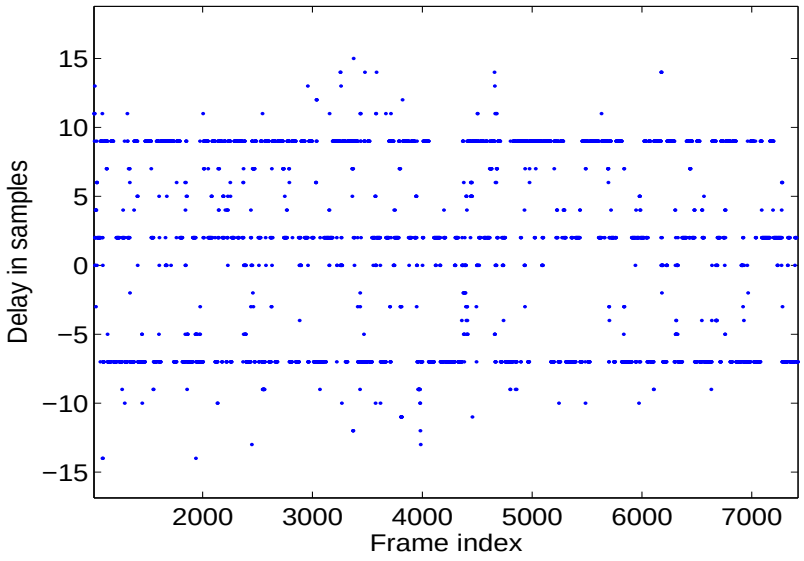

Fig. 3. Location of peaks in samples as a function of frame index.

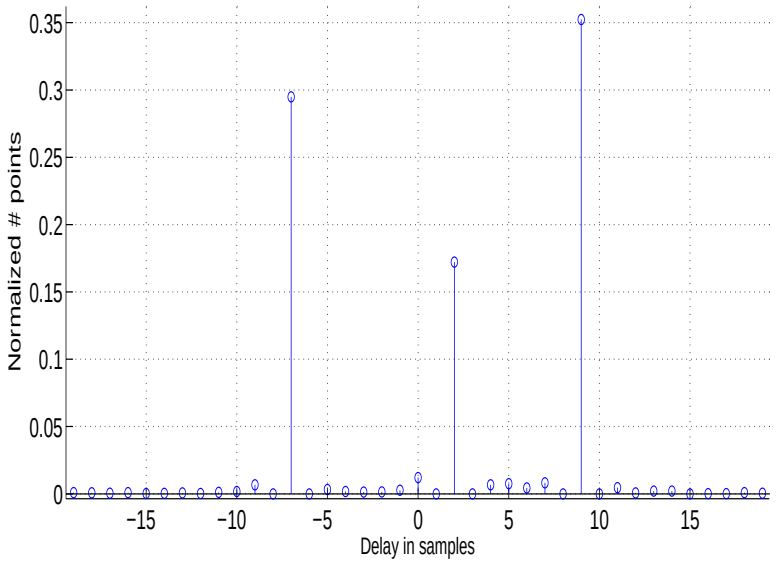

Fig. 4. Number of points for each delay normalized with respect to the total number of points.

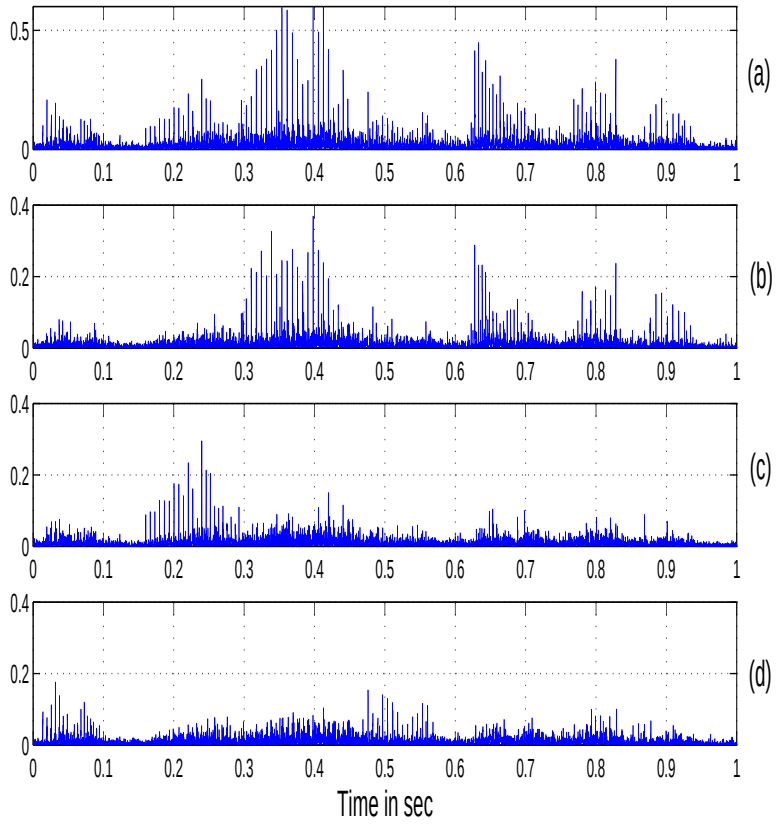

Fig. 5. (a) HE of *mic-1* signal, (b) HE of *speaker-1*, (c) HE of *speaker-2* and (d) HE of *speaker-3*.

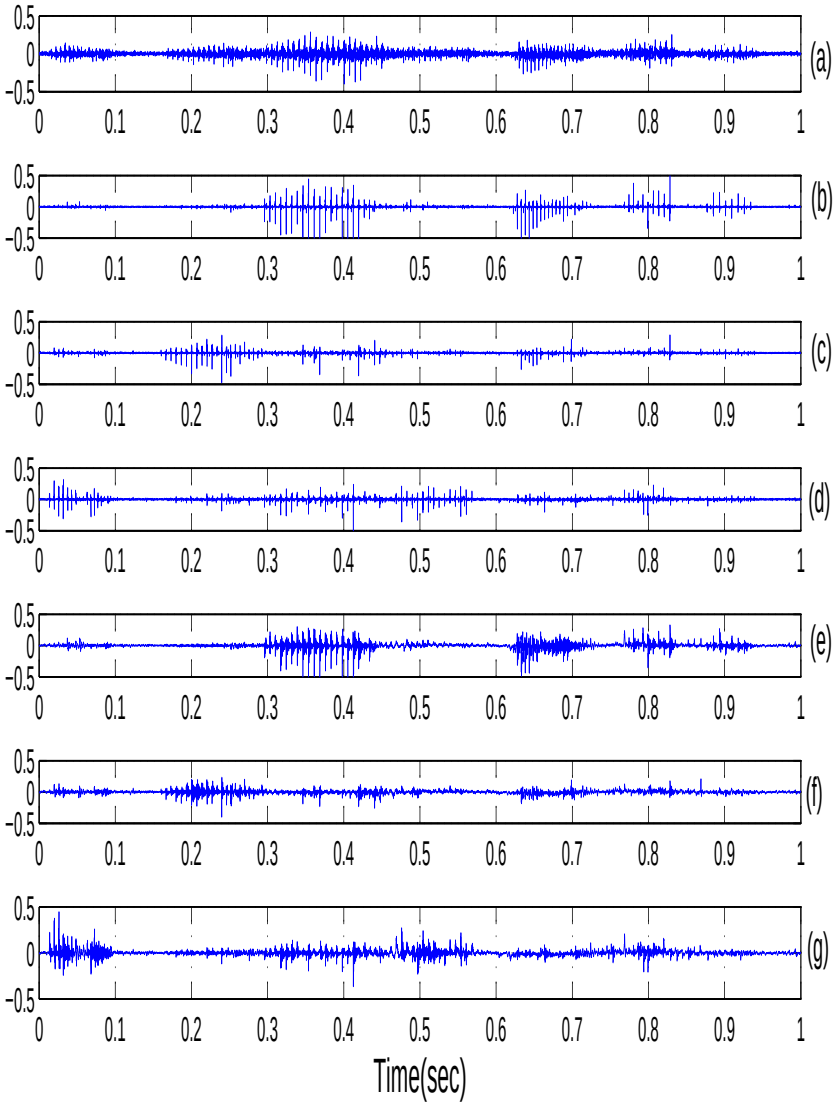

Fig. 6. (a) LP residual of *mic-1* signal,(b), (c) and (d) Modified residuals of *speaker1, speaker2* and *speaker3* respectively, (e), (f) and (g) Synthesized speech signal of*speaker1, speaker2*and *speaker3* respectively.

4 Summary and conclusions

In this paper, we have proposed a method to separate speech signals of individual speakers from multispeaker speech signal using excitation source information. The results show that it is indeed possible to separate the speech of individual speakers from a two microphone data. The separation would be better if the data from the *mic-2* is also used for separation of the signals. Due to significant differences in the levels of speech of the speakers, it is not possible to separate the speakers effectively using only two microphone data. If more number of spatially distributed microphones are used, then it is possible to separate the speakers significantly better than what can be achieved from two microphones. In practice there may be some movement of the speakers. In such a case the delays have to be computed as a function of time. These issues are being addressed in our ongoing research.

References

1. O.M.M.Mitchell, C.A.Ross, G.H.Yates: Signal processing for a cocktail party effect. J. Acoust. Soc. Amer. **50** (1971) 656–660
2. D.P.Morgan, E.B.George, L.T.Lee, S.M.Kay: Cochannel speech separation by harmonic enhancement and supression. IEEE Trans. Speech Audio Processing **5** (1997) 407–424
3. A.K.Barros, T.Rutkowski, F.Itakura, N.Ohnishi: Estimation of speech embedded in a reverberant and noisy environment by independent component analysis and wavlets. IEEE Trans. Neural Networks **13** (2002) 888–893
4. B.Yegnanarayana, S.R.M.Prasanna, M.Mathew: Enhancement of speech in multispeaker environment. In: Proc. European Conf. Speech Processing, Technology, Geneva, Switzerland (2003) 581–584
5. Cardoso, J.F.: Blind signal separation: Statistical principles. Proc. IEEE **86** (1998) 2009–2025
6. F.Asano, S.Ikeda, M.Ogawa, H.Asoh: Combined approach of array processing and independent component analysis for blind separation of acoustic signals. IEEE Trans. Speech Audio Processing **11** (2003.) 204–215
7. S.C.Douglas, H.Sawada, S.Makino: Natural gradient multichannel blind deconvolution and speech separation using causal fir filters. IEEE Trans. Speech Audio Processing **13** (2005) 92–104
8. B.Yegnanarayana, Prasanna, S., R.Duraiswamy, D.Zotkin: Processing reverberant speech for time-delay estimation. Accepted for publication in IEEE Trans. Speech Audio Processing (2004)
9. V.C.Raykar, B.Yegnanarayana, Prasanna, S., R.Duraiswamy: Speaker localization using excitation source information in speech. Accepted for publication in IEEE Trans. Speech Audio Processing (2004)
10. B.Yegnanarayana, Prasanna, S.: Two speaker speech enhancement using excitation source information. submitted to Journal of Accoustic Society of America (under review) (2004)
11. J.Makhoul: Linear prediction: A tutorial review. Proc. IEEE **63** (1975) 561–580
12. V.Oppenheim, A., W.Schafer, R.: Digital signal processing. Prentice Hall, Englewood Cliffs, New Jersey (1975)
13. T.V.Ananthapadmanabha, B.Yegnanarayana: Epoch extraction from linear prediction residual for identification of closed glottis interval. IEEE Trans. Acoust., Speech, Signal Processing **ASSP-27** (1979) 309–319