



Time-Delay Estimation Using Source and Spectral Information from Speech

S. R. Mahadeva Prasanna, Kiran Bakki and P. Krishnamoorthy
Department of Electronics and Communication Engineering
Indian Institute of Technology Guwahati
Guwahati-781039, Assam, India
{prasanna, k.bakki, pkm}@iitg.ernet.in

Abstract—This paper proposes a method for estimating the time-delay between speech signals collected over two spatially distributed microphones using source and spectral information. For a given frame, the cross-correlation sequences obtained separately using the excitation source and spectral information are combined and the time-delay is estimated. The time-delay estimated from the combined cross-correlation sequence is found to be providing better performance compared to the time-delay estimated using individual cross-correlation sequences from source and spectral information. The improved performance by the proposed combined method demonstrates the complementary nature of the information exploited by the source and spectral based methods for time-delay estimation.

I. INTRODUCTION

Time-delay estimation is an important and integral component of multi-channel signal processing [1]. Multi-channel signal processing is found to provide better performance compared to single channel processing in many applications like speech enhancement, speech and speaker recognition, multi-speaker processing, speaker localization and tracking [2]. The main idea behind the multi-channel signal processing is the availability of multiple signals simultaneously collected over spatially distributed sensors, which helps in enhancing the required component and suppressing all other unwanted degradation components. This is conveniently achieved by first estimating the time-delay between multi-channel signals and using the knowledge of estimated time-delay for further processing [2]. Thus estimation of time-delay is an important research issue in multi-channel signal processing. This paper proposes a robust method for time-delay estimation in case of multi-channel signal processing by exploiting the excitation source and spectral information.

Time-delay estimation by the cross-correlation of signals from two channels is the most convenient approach [3]–[5]. The Generalized Cross-Correlation (GCC) approach is the mostly used approach for time-delay estimation [3]. The GCC approach is suitable for any type of multi-channel signal, including speech. However, recently few methods have been proposed which are tailored for time-delay estimation in case of speech [6]–[8]. In [6] the authors exploit the harmonic nature of the speech signal for time-delay estimation. In [7], [8] the authors exploit the nature of the excitation source for time-delay estimation.

The GCC approach is a spectral based approach whose performance is dependent on how best the spectral information of wanted signal (speech) is manifested and also on how best the spectral information due to degradation (noise and reverberation) is nullified. Alternatively, in the time-delay estimation based on the excitation source information the performance depends on the impulsive nature of excitation source [7]. The GCC approach is generally known to give better performance when blocks for cross-correlation are considered for long duration typically 1-3 sec. However, the excitation source based method can work even for short segments of speech like 50-100 ms [7]. Any cross-correlation based approach works well against background noise. The excitation source based approach is also a cross-correlation based approach, but it is shown to perform better compared to GCC approach, even in reverberant environment and also when the block size is as short as 50-100 ms. Smaller block sizes are useful in applications like speaker localization and tracking. As described above the GCC based approach exploits the spectral information and the excitation source information exploits the impulsive nature of excitation source for time delay estimation. Thus the two methods may be combined to obtain a robust method for time-delay estimation, especially in case of smaller block sizes of 50-100 ms.

In the present work the speech is processed by the Linear Prediction (LP) analysis to extract the LP residual signal which mostly contains excitation information [9]. However, the magnitude of the analytic signal of the LP residual termed as Hilbert envelope is known to have better manifestation of impulsive nature of excitation information [10]. Hence, the Hilbert envelope of the LP residual is used in this study as the excitation information [7]. The Hilbert envelopes of the two microphone signals are processed in blocks of 50 ms with a shift of 10 ms to compute the cross-correlation sequence. In a similar way the speech signals from the two microphones are processed in blocks of 50 ms with a shift of 10 ms by the spectral based GCC approach. For each block of 50 ms, the two cross-correlation sequences are combined and the time-delay is estimated from the combined cross-correlation sequence.

The rest of the paper is organized as follows: Section 2 explains the basis of excitation source information based



time-delay estimation method. Section 3 describes the spectral based approach for time-delay estimation. The proposed combined source and spectral information based time-delay estimation method is described in Section 4. Section 5 gives the experimental results and discussion related to the same. The summary of the present work and the scope for the future work are given in section 6.

II. TIME-DELAY ESTIMATION USING EXCITATION SOURCE INFORMATION

Let $s_1(n)$ and $s_2(n)$ be the speech signals collected over two spatially distributed microphones mic-1 and mic-2. The speech signals are processed by the LP analysis to extract the LP residual signals [9]. This is achieved by estimating the predictable components in terms of the linear prediction coefficients (LPCs), constructing the inverse filter using the LPCs and then passing the speech signal through the inverse filter [9]. Thus the LP residual of mic-1 signal is computed as $r_1(n) = s_1(n) * h_1(n)$, where $h_1(n) \leftrightarrow H_1(z)$, $H_1(z) = 1 + \sum_{k=1}^p a_k z^{-k}$ and a_k are the LPCs estimated from $s_1(n)$ by LP analysis. Similarly, the LP residual of mic-2, that is, $r_2(n)$ is computed.

The Hilbert transform (HT) of $r_1(n)$ is given as

$$rh_1(n) = \begin{cases} IDFT(jR_1(\omega)) & \text{for } -\Pi < \omega \leq 0 \\ IDFT(-jR_1(\omega)) & \text{for } 0 < \omega \leq \Pi \end{cases} \quad (1)$$

where $R_1(\omega)$ is the Discrete Fourier Transform (DFT) of $r_1(n)$ and IDFT indicates the Inverse Discrete Fourier Transform. Similarly the Hilbert transform of $r_2(n)$, that is, $rh_2(n)$ is computed. The Hilbert envelope of the LP residual $r_1(n)$ is given by $he_1(n) = \sqrt{r_1^2(n) + rh_1^2(n)}$. Similarly, the Hilbert envelope of the LP residual $r_2(n)$, that is, $he_2(n)$ is computed. The cross-correlation sequence of the two Hilbert envelopes is given by

$$C_h(n) = he_1(n) * he_2(n) \quad (2)$$

The speech signals from the two microphones, their LP residuals and the Hilbert envelopes are given in Fig. 1. The cross-correlation sequence of the two Hilbert envelopes is given in Fig. 2 and the shift in the peak from the centre of the cross-correlated sequence gives the time-delay between the two microphone signals.

III. TIME-DELAY ESTIMATION USING SPECTRAL INFORMATION

Let $s_1(t)$ and $s_2(t)$ be the speech signals collected over two spatially distributed microphones mic-1 and mic-2. The time-delay is estimated by computing the cross-correlation between the speech signals obtained by the two microphones. The time shift causing the peak gives the estimate of the delay. The cross-correlation between two signals $s_1(t)$ and $s_2(t)$ is given by

$$C_g(t) = \int_{-\infty}^{\infty} s_1(\tau)s_2(t - \tau)d\tau \quad (3)$$

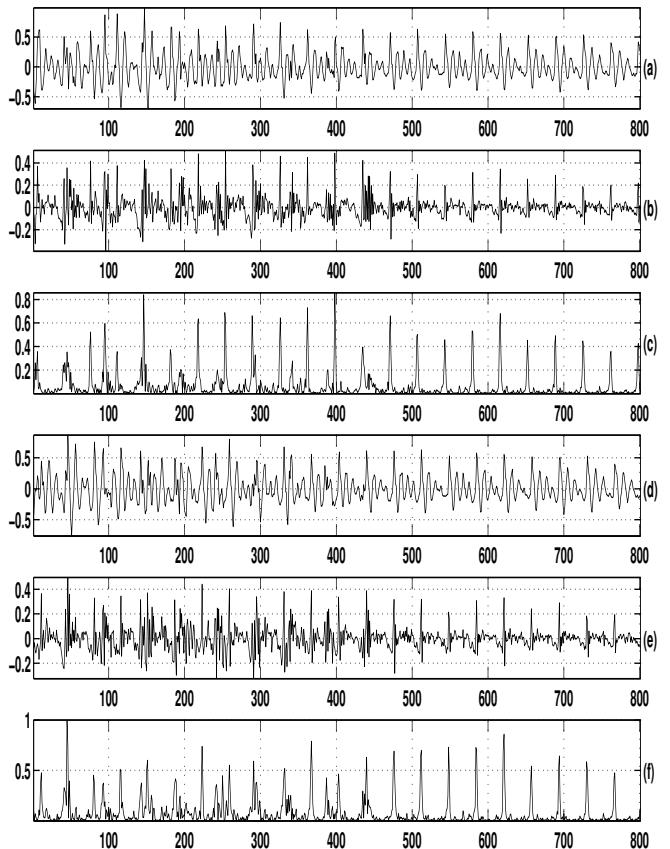


Fig. 1. Nature of the Hilbert Envelope of LP residual: (a) Speech signal collected, b LP residual and (c) Hilbert envelope of the LP residual for the signal at mic-1. (d) Speech signal, (e) LP residual, and (f) Hilbert envelope of the LP residual for the signal at mic-2.

The cross-correlation of $s_1(t)$ and $s_2(t)$ is related to the cross power spectral density (PSD) by the Fourier transform relation

$$C_g(t) = \int_{-\alpha}^{\alpha} G_{s_1 s_2}(f) e^{j2\Pi f \tau} df \quad (4)$$

where $G_{s_1 s_2}(f)$ is the cross PSD of $s_1(t)$ and $s_2(t)$.

In order to improve the accuracy of delay estimate, it is desirable to pre-filter $s_1(t)$ and $s_2(t)$. If $s_1(t)$ is filtered through $H_1(f)$ and $s_2(t)$ is filtered through $H_2(f)$, then the cross-correlation of $s_1(t)$ and $s_2(t)$ after pre-filtering becomes

$$C_g(t) = \int_{-\alpha}^{\alpha} H_1(f)H_2^*(f)G_{s_1 s_2}(f)e^{j2\Pi f \tau} df \quad (5)$$

$$C_g(t) = \int_{-\alpha}^{\alpha} \Psi_g(f)G_{s_1 s_2}(f)e^{j2\Pi f \tau} df \quad (6)$$

where $\Psi_g(f) = H_1(f)H_2^*(f)$ denotes general frequency weighting [3]. The weight function is chosen to ensure a large sharp peak rather than a broad one in order to ensure good

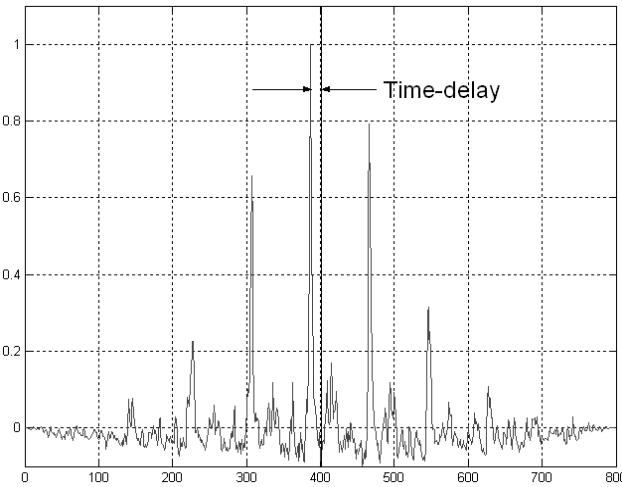


Fig. 2. The cross-correlation sequence of the Hilbert envelopes of size 50 ms (400 samples) for the signals at mic-1 and mic-2. The time-delay is obtained as 14 samples.

time-delay resolution. However, sharp peaks are sensitive to errors introduced by the finite observation time, particularly in cases of low SNR. Thus the choice of $\Psi_g(f)$ is a compromise between good resolution and stability. The phase transform is most commonly used as weight function. The weighting function used in our work is the phase transform and is given as [3]

$$\psi_g(f) = 1/|G_{s_1 s_2}(f)| \quad (7)$$

The cross-correlation between the two microphone signals using phase transform as weighting function is given as [3]

$$C_g(t) = \int_{-\infty}^{\infty} \{G_{s_1 s_2}(f)/|G_{s_1 s_2}(f)|\} e^{j2\Pi f \tau} \quad (8)$$

The cross-correlated sequence of the two microphone signals is shown in Fig. 3 and the shift in the peaks from the centre of the cross-correlated sequence gives the time-delay between the two microphone signals.

IV. TIME-DELAY ESTIMATION USING SOURCE AND SPECTRAL INFORMATION

In section 2 we discussed about the basic principles behind the source feature based time-delay estimation method proposed in [7]. In section 3 we discussed about the basic principles behind the spectral feature based time-delay estimation method proposed in [3].

From this discussion it is evident that even though both the methods employ cross-correlation, they exploit different features from the speech signal for the time-delay estimation. The excitation source feature based method exploits the impulsive nature of the excitation sequence and spectral feature based method exploits the spectral information. This observation motivated us to explore the possibility of combining these two methods for time-delay estimation. It was observed that

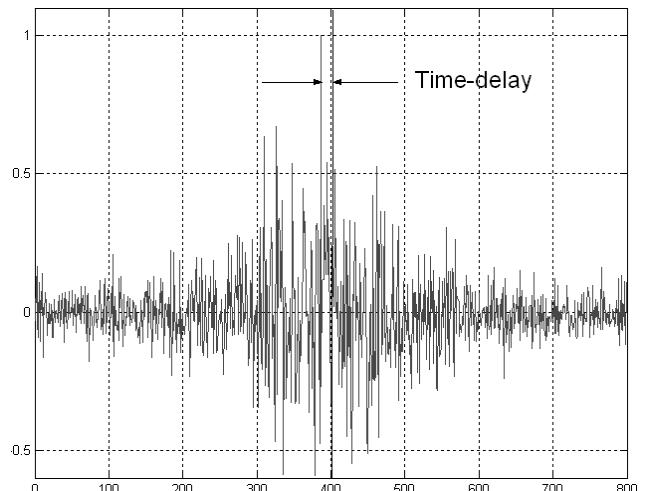


Fig. 3. The cross-correlation sequence of the two microphone signals of size 50 ms (400 samples) using phase transform as weighting function.

the wrong time-delay estimation in either source based or spectral based GCC method was due to the spurious peak being larger than the true peak corresponding to the time-delay. Since the two methods exploit different information from the speech signal, it is unlikely that the cross-correlation sequence of both the methods will have the spurious peak at the same instant. Thus by considering a minimum from the two cross-correlation sequences, we can obtain a combined cross-correlation sequence in which the spurious peak is eliminated. This is illustrated in Fig. 4. As it can be observed in Fig. 4(b) the cross-correlation sequence by the GCC approach contains a spurious peak which results in wrong estimation of time-delay. However it can be observed from the cross-correlation sequence for the same frame using the excitation source feature based method does not contain the spurious peak. The combined cross correlation sequence is given by

$$C_c(n) = \min\{C_h(n), C_g(n)\} \quad (9)$$

and is shown in Fig. 4(c). The sample minima computed in Eqn. (9) of the two cross-correlation sequence helps in reducing the effect of spurious peaks, while preserving the genuine one. The excitation The time-delay estimation from the combined cross-correlation sequence minimizes the spurious delays. The same is true for some other frames where the cross-correlation sequence by the excitation source based method contains spurious peak and the GCC method does not contain the spurious peak.

V. EXPERIMENTAL RESULTS AND DISCUSSION

The speech signals are collected over two spatially distributed microphones in a laboratory environment of dimension $5.67 \times 4.53 \times 2.68$ m, which contained some computers and partitions. The two microphones are separated by 0.6 m, and they are about 1 m distance from the speakers. The speech signals are sampled at 8 kHz and stored as 16 bit samples. The

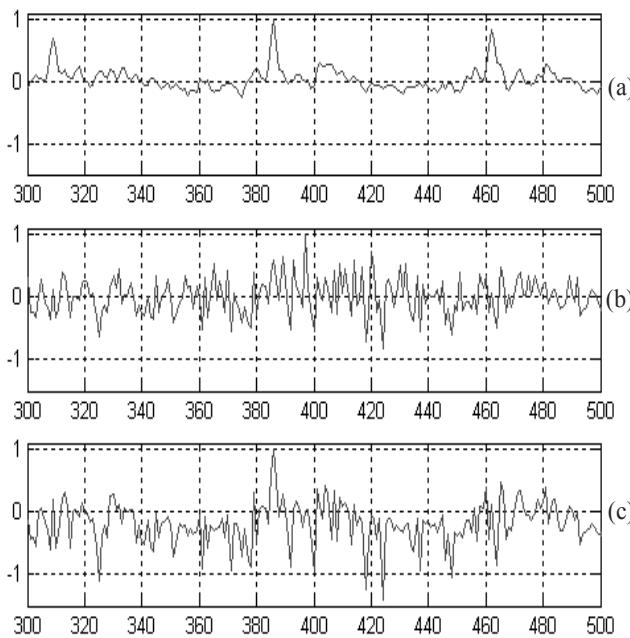


Fig. 4. Segments of cross-correlation sequence by (a) Excitation source method giving correct delay, (b) Spectral method giving the spurious delay and (c) combined method giving correct delay.

speech signals are processed by the LP analysis using a block size of 20 ms and block shift of 10 ms using 10th order LP analysis to extract LP residuals and their Hilbert envelopes. The signals are processed by the excitation source feature based and spectral based methods in blocks of 50 ms with a shift of 10 ms. The time-delays estimated by these methods are shown in Fig. 5. Similarly the time-delays estimated by the proposed combined is also shown in Fig 5. As it can be observed the number of spurious delays is less in the combined method compared to the individual methods.

An objective measure for comparison of the performance could be the ratio (r) of the number of points around the time-delay within ± 1 sample deviation to the total number of points above a certain threshold of the value of the standard deviation of the samples of the Hilbert envelope. Since lower values of the standard deviation correspond mostly to the unvoiced region, we can ignore the values below 0.25 for computing this ratio. The values of r for different cases are shown in Fig. 5. The larger the value of r , the better is the method for estimating the time-delay. From these illustrations we can infer that the proposed combined method provides superior performance.

VI. SUMMARY AND CONCLUSIONS

In this paper we have proposed a method for time-delay estimation using source and spectral information. The proposed method is based on the fact that the source based and spectral based methods exploit different information from the speech signals for time delay estimation. The performance of the proposed method is found to be better compared to

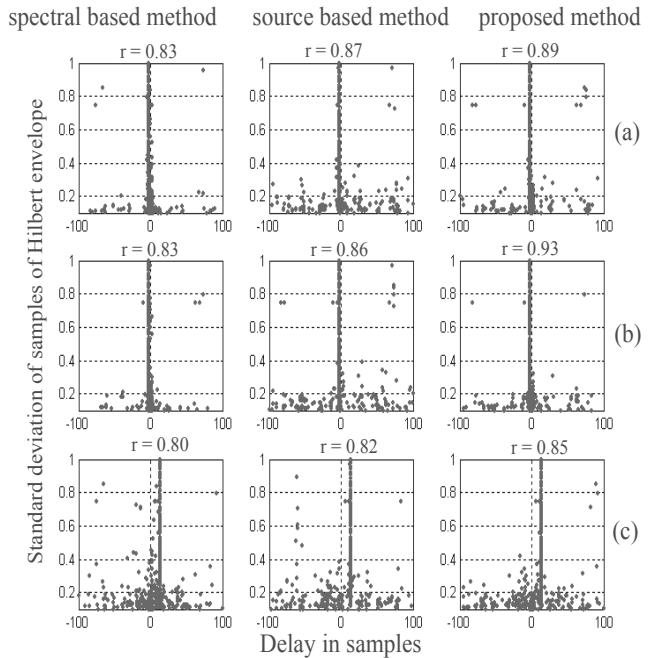


Fig. 5. Performance Evaluation: (a) mic-1 and mic-2 signals, (b) mic-1 and mic-3 signals, and (c) mic-1 and mic-4 signals.

individual methods. A method for speaker localization and tracking can be developed using the proposed time-delay estimation method. Further, the effect multipath environment in the proposed method needs to be studied.

REFERENCES

- [1] J. Chen, J. Benesty, and Y. A. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP J. Applied Signal process.*, vol. 2006, Article ID 26503, 19 pages, doi:10.1155/ASP/2006/26503.
- [2] S. R. M. Prasanna, "Event based analysis of speech," Ph.D. dissertation, Indian Institute of Technology Madras, Dept. of Computer Science and Engg., Chennai, India, Mar. 2004.
- [3] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech and Signal process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.
- [4] D. Hertz, "Time delay estimation by combining efficient algorithms and generalized cross-correlation methods," *IEEE Trans. Acoustics, Speech and Signal process.*, vol. 34, no. 1, pp. 1–7, Feb. 1986.
- [5] B. Chen and P. Loizou, "Speech enhancement using a MMSE short time spectral amplitude estimator with Laplacian speech modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal process.*, vol. 1, Philadelphia, PA, USA, 2005, pp. 1097–1100.
- [6] M. S. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *J. Acoust. Soc. Am.*, vol. 105, pp. 2914–2919, May 1999.
- [7] B. Yegnanarayana, S. R. M. Prasanna, R. Duraiswami, and D. Zotkin, "Processing of reverberant speech for time-delay estimation," *IEEE Trans. Speech Audio process.*, vol. 13, pp. 1110–1118, Nov. 2005.
- [8] V. C. Raykar, B. Yegnanarayana, S. R. M. Prasanna, and R. Duraiswami, "Speaker localization using excitation source information in speech," *IEEE Trans. Speech Audio process.*, vol. 13, no. 5, pp. 751–761, Sep. 2005.
- [9] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.
- [10] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal process.*, vol. ASSP-27, pp. 309–319, Aug. 1979.