Tracking A Moving Speaker using Excitation Source Information

Vikas C. Raykar, Ramani Duraiswami

Perceptual Interfaces & Reality Laboratory,
Institute of Advanced Computer Studies,
University of Maryland,
College Park, MD 20742
Email:{vikas,ramani}@umiacs.umd.edu

B. Yegnanarayana, S.R. Mahadeva Prasanna

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

Abstract

Microphone arrays are widely used to detect, locate, and track a stationary or moving speaker. The first step is to estimate the time delay, between the speech signals received by a pair of microphones. Conventional methods like generalized cross-correlation are based on the spectral content of the vocal tract system in the speech signal. The spectral content of the speech signal is affected due to degradations in the speech signal caused by noise and reverberation. However, features corresponding to the excitation source of speech are less affected by such degradations. This paper proposes a novel method to estimate the time delays using the excitation source information in speech. The estimated delays are used to get the position of the moving speaker. The proposed method is compared with the spectrum-based approach using real data from a microphone array setup.

1. Introduction

Many applications require the capture of high quality speech information from users who are not tethered to a close speaking microphone [1,2]. In such conditions locating and tracking the speaker in the acoustical environment is essential for effective communication. For instance, tracking a moving speaker is important in applications such as video-conferencing or meeting or lecture summarization, where the speaker may be moving continuously. In this case, information about the moving speaker can be obtained from the speech signal. This information can then be fed to a video system for actuating camera pan-tilt operations to keep the speaker in focus automatically [3,4]. This provides a significant improvement in the overall effect of audio-visual communication for the far-end listeners. Tracking a moving speaker is also useful in multispeaker processing in which speech from a particular speaker may be enhanced with respect to others, or with respect to noise sources.

The speech signal received from a speaker in an acoustical environment is corrupted both by additive noise as well as room reverberation. In the case of a moving speaker, this is further complicated by the change in the characteristics of reverberation, as the speaker moves from one place to the other, due to the variability of the room impulse response with both source and receiver locations. One effective way of handling such a situation is the use of a set of spatially distributed microphones for recording the speech. The signal received by several microphones is processed to obtain information about the time-delay between pairs of microphones. The estimated time-delays for pairs of microphones can be used for computing location of the speaker, which can then be used for tracking.

Most of the methods for time delay estimation are based on finding the time lag which maximizes the cross-correlation between filtered versions of the received signals. The most commonly used method is the Generalized Cross Correlation method proposed by Knapp and Carter [5]. The GCC function $R_{x_1x_2}(\tau)$ is computed as [5]

$$R_{x_1x_2}(\tau) = \int_{-\infty}^{\infty} W(\omega) X_1(\omega) X_2^*(\omega) e^{j\omega\tau} d\omega \qquad (1)$$

where $X_1(\omega)$, $X_2(\omega)$ are the Fourier transforms of the microphone signals $x_1(t)$, $x_2(t)$, respectively and $W(\omega)$ is the weighting function. The two commonly used weighting functions are the Phase Transform (PHAT) and the Maximum likelihood (ML) weighting [5]. This ML weighting function performs well for low room reverberation. As the room reverberation increases this method shows severe performance degradations [6]. The PHAT weighting $W_{PHAT}(\omega)$ is the other extreme where we completely flatten out the magnitude spectrum and is given by $W_{PHAT}(\omega) = 1/|X_1(\omega)X_2^*(\omega)|$. By flattening out the magnitude spectrum the resulting peak in the GCC function corresponds to the dominant delay. However, the disadvantage is that it works well only when the noise level is low. All these methods do not exploit the mechanism of speech production to get robust estimates. Recently, Brandstein [7] proposed a method based on the explicit knowledge of the periodicity of voiced speech.

However, most of the existing methods use the spectral features which mostly correspond to the vocal tract system information in case of speech. The spectral features are corrupted during transmission due to the medium, noise and the room reverberation. However, we show that the features corresponding to excitation source information are robust to such degradations. We discuss methods to extract the excitation source information from the speech signal and use this to estimate the time delay.

The paper is organized as follows: In Section 2 a method for time-delay estimation using the excitation source information is discussed. A method for tracking a moving speaker using the estimated delays from the excitation source information is proposed in Section 3. Section 4 describes experimental results, as well as comparison with a spectral-based GCC-PHAT approach. The paper concludes with a summary of the present work, and with a discussion on possible extensions.

2. Time-Delay Estimation using Excitation Source Information

Speech is the result of excitation of a time-varying vocal tract system with time-varying excitation [8]. The common and significant mode of excitation of the vocal tract system is the voiced

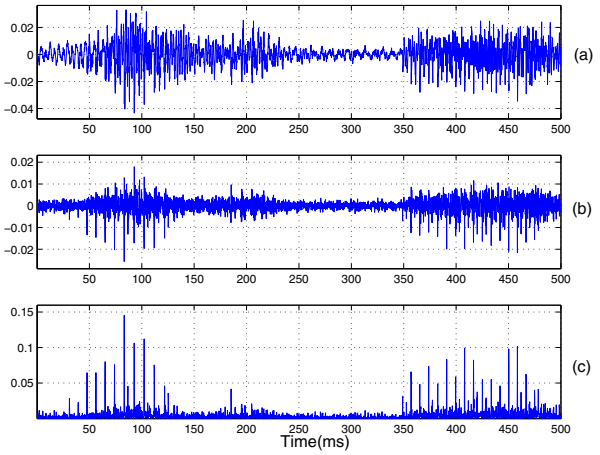

Figure 1: (a) Voiced segment ($s(n)$), (b) 10^{th} order LP residual ($r(n)$), and (c) Hilbert envelope of the LP residual($h(n)$).

excitation caused by the vibrating vocal folds at the glottis, which to a first approximation may be treated as consisting of a sequence of impulses. The vocal tract system information is represented in terms of spectral features, which may be assumed to be superimposed on the glottal excitation pulses. The spectral features due to the vocal tract get corrupted due to the transmission medium, noise and the room response. However the location of the epochs i.e. the instants of significant excitation are not affected by the transfer characteristics of the microphones and the medium.

The excitation source information from the given speech signal can be extracted by using the linear prediction (LP) analysis [9]. In the LP analysis each sample is predicted as a linear combination of the past p samples, where p is the order of prediction. If $s(n)$ is the given speech sequence, then its predicted value is given by,

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{2}$$

where $\{a_k\}$ are the LP coefficients. The error between the given speech sequence and that of its predicted one is given by

$$r(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \tag{3}$$

This error is termed as LP residual. From the given speech signal, the LP residual can be extracted by passing it through an inverse filter given by

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

In the LP residual most of the spectral envelope information is removed. So the spectral degradations due to noise and reverberation are eliminated. Since the locations of epochs are robust to degradations, the peaks in the cross-correlation of LP residuals are due to the epochs.

However the LP residual signal amplitude fluctuates depending on the phase of the signal. Hence if we directly use to LP residuals, it may result in a poor correlation peak. Therefore, instead of using the LP residual directly, a new feature called the Hilbert envelope of the LP residual is used, which is

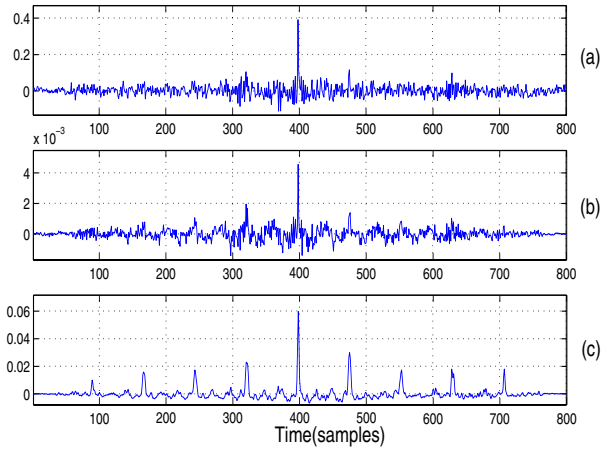

Figure 2: (a) GCC with PHAT weighting, (b) Cross-correlation of the 10^{th} order LP residuals and (c) Cross-correlation of the Hilbert envelopes of the corresponding LP residuals of two 50 ms speech segments

defined as [10, 11]

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

where $r_h(n)$ is the Hilbert transform of $r(n)$ and is computed by passing $r(n)$ through a filter whose response is given by

$$H(f) = -j\ sgn(f) \tag{6}$$

Figure 1 shows a segment of voiced speech captured in a noisy reverberant environment, the 10^{th} order LP residual and the corresponding Hilbert envelope of the LP residual. The peaks in the Hilbert envelope correspond to the locations of the epochs.

The time-delay between a pair of microphones is estimated by computing the cross-correlation function of the Hilbert envelopes of the LP residuals. For every frame, the cross correlation function is computed, and the displacement of the peak with respect to the center is noted as the time-delay. Figure 2(a) shows the Generalized cross-correlation function with PHAT weighting, between two 50 ms speech segments recorded in a noisy reverberant room. Figure 2(b) shows the cross-correlation sequence for the 10^{th} order LP residual of the two speech segments. The plot looks similar to that of the GCC case. Figure 2(c) shows the cross-correlation sequence for the Hilbert envelopes of the LP residuals. As can be seen Figure 2(c) shows a significantly prominent peak (with respect to the samples surrounding it) compared to the previous two cases. The reason for this is that, in the Hilbert envelopes of the LP residuals, the high SNR portions correspond to the major excitations (epochs) of the vocal tract system. The high SNR excitation information at the epochs is preserved better in the Hilbert envelope, than in the speech signal or in its LP residual. These high amplitude values at the epochs dominate in the computation of the cross-correlation sequence.

3. Tracking A Moving Speaker

Once the time delays are estimated the source localization problem can be formulated as follows: Let there be M pairs of microphones. Let $\mathbf{m_i^1}$ and $\mathbf{m_i^2}$ for $i \in [1, M]$ be the vectors representing the spatial coordinates of the two microphones in the i^{th} pair of microphones. Let the source be located at $\mathbf{s}$. The

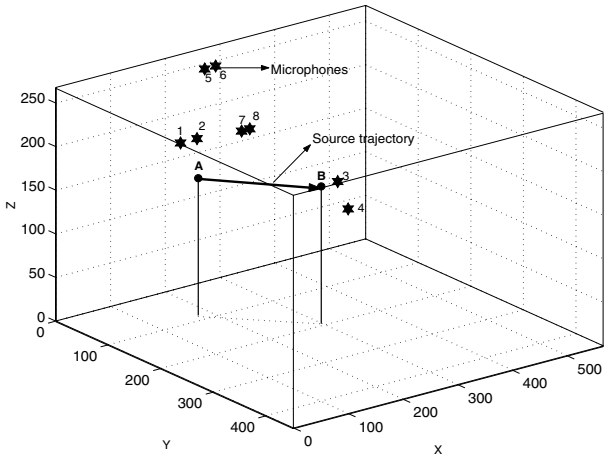

Figure 3: Schematic of the microphone array setup and the source trajectory.

actual delay associated with a source at $\mathbf{s}$ and the i^{th} pair of microphones is given by,

$$t_i(\mathbf{s}) = \frac{|\mathbf{s} - \mathbf{m_i^1}| - |\mathbf{s} - \mathbf{m_i^2}|}{c} \tag{7}$$

where, c is the speed of propagation of sound in the acoustical medium. In practice, for a given microphone pair, the estimated delay τ_i and the actual delay $T_i(\mathbf{s})$ will never be equal because the estimated delay is corrupted by noise.

Let τ_i be the estimated time-delay. Let τ_i the estimated time delay be corrupted by zero-mean additive white Gaussian noise with known variance $var(\tau_i)$. So τ_i is normally distributed with mean $t_i(\mathbf{s})$ and variance $var(\tau_i)$.

$$\tau_i \sim \mathrm{N}(t_i(\mathbf{s}), var(\tau_i)) \tag{8}$$

Assuming that each of the time delays are independently corrupted by zero-mean additive white Gaussian noise the likelihood function can be written as:

$$p(\tau_1, \tau_2, ..., \tau_M; s) = \prod_{i=1}^{M} \frac{1}{\sqrt{2\pi var(\tau_i)}} exp[\frac{-(\tau_i - t_i(\mathbf{s}))^2}{2var(\tau_i)}] \tag{9}$$

The log-likelihood ratio is:

$$ln(p(\tau_1, \tau_1, ..., \tau_M; s)) = -\sum_{i=1}^{M} ln(\sqrt{2\pi var(\tau_i)}) + [\frac{(\tau_i - t_i(\mathbf{s}))^2}{2var(\tau_i)}] \tag{10}$$

The Maximum Likelihood(ML) location estimate, $\hat{\mathbf{s}}_{ML}$is the position which maximizes the log likelihood ratio or equivalently one which minimizes:

$$J_{ML}(\mathbf{s}) = \sum_{i=1}^{M} \frac{[\tau_i - t_i(\mathbf{s})]^2}{var(\tau_i)} \tag{11}$$

$$\hat{\mathbf{s}}_{ML} = \arg_{\mathbf{s}}(min(J_{ML}(\mathbf{s}))) \tag{12}$$

This does not have a closed-form solution since it is a non-linear function of s. In our experiments we used the Gauss-Newton nonlinear least square method [12] to minimize this function.

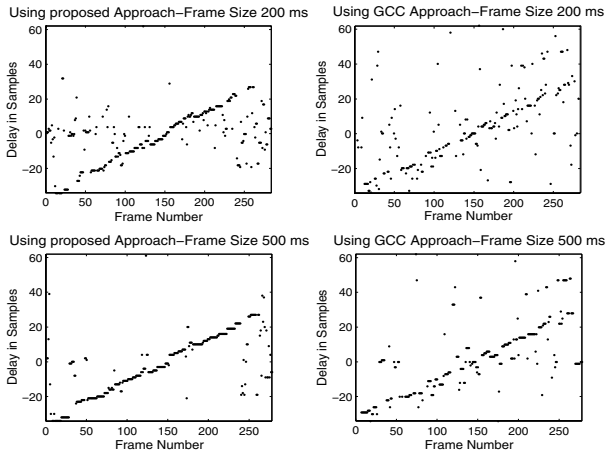

Figure 4: Estimated delay as a function of frame number for one microphone pair(*MIC-1* and *MIC-4*) using the proposed approach and the GCC approach for frame length $200ms$ and $500ms$ with frame shift of $50ms$.

4. Experimental Results

The experimental setup consisted of a 14 element electret microphone array with additional hardware to capture the data into the computer, placed in a normal room of dimension $5.67 \times 4.53 \times 2.68m$, whose reverberation time was approximately $200ms$. The room had 5 computers which contributed to the ambient noise in the room. The room also had partitions which contributed to some reflections in the room. Out of the 14 microphones, only 8 (say, *MIC-1* to *MIC-8*) were used in the present study. Figure 3 shows the schematic of the room and the position of the microphones. Each microphone data was sampled at 8 kHz, and stored with 16 bit resolution. For all the experiments the speaker was instructed to move in the room reading a text at his normal level of speaking. In order to validate the results, the speaker was asked to move in a predetermined path whose coordinates were known. We performed four set of experiments corresponding to different paths. Due to space constraints we will show the results for only one case.

In one set of experimentation the speaker moved from one end of the room towards the microphone array along the trajectory as shown in Figure 3. Figure 4 shows the estimated time-delays for one pair of microphones (*MIC-1 & MIC-4*), for the proposed approach and the generalized cross-correlation(GCC) approach [5] with PHAT weighting, for every frame of $200ms$ and $500ms$ with a shift of $50ms$. The delays estimated by the proposed approach are more uniform compared to those estimated using the GCC-PHAT approach. Also as the frame length is increased the delays obtained are much more consistent. Figure 5 and 6 shows the actual and estimated x, y and z coordinates of the speaker for every frame of $200ms$ and $500ms$ respectively. The actual source trajectory is shown in solid line while the estimated is shown in dots. It can be seen that the estimated source trajectory follows the actual trajectory closely for the proposed approach than the GCC approach. Figure 7 shows the corresponding *localization error* as a function of the frame number. The *localization error* is defined as the the Euclidean distance between the actual position and the estimated position. It can be seen that on an average the *localization error* for the proposed method is less than that of the GCC-PHAT method. Similar results were observed for all the experiments conducted.

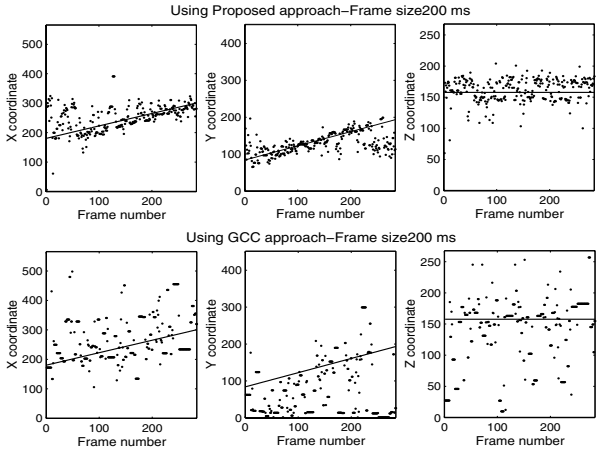

Figure 5: The actual and the estimated x,y and z coordinates of the speaker using the proposed approach and using the GCC approach for every frame $200ms$ with a shift of $50ms$. The actual path is shown with solid line and the estimated path is shown with dots.

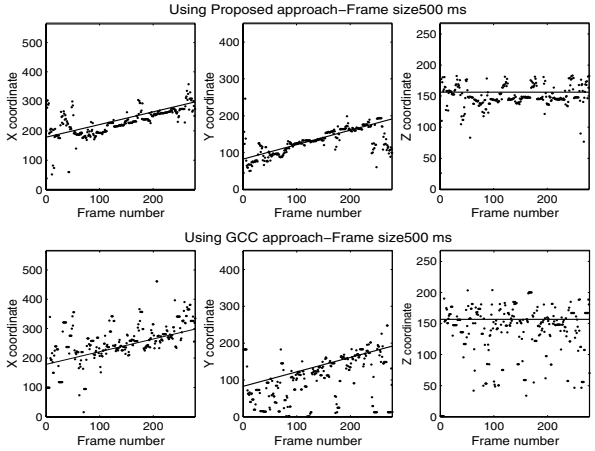

Figure 6: The actual and the estimated x,y and z coordinates of the speaker using the proposed approach and using the GCC approach for every frame of $500ms$ with a shift of $50ms$. The actual path is shown with solid line and the estimated path is shown with dots.

5. Conclusions

In this paper a method for estimation of time-delays and tracking a moving speaker using the excitation source information in the speech signal is discussed. Comparison of the results obtained from the proposed approach with that of the existing spectral-based approach (GCC-PHAT) show that the parameters estimated by the proposed approach are more closer to the actual values. Both the vocal tract system features as well as the excitation source features contain significant information about the moving speaker. The potential of the vocal tract system features has already been established. In this paper usefulness of the excitation source information is illustrated. An effective way of combining these two approaches may result in a robust estimation of various parameters required for tracking a moving speaker.

6. References

[1] S. Oh and V. Viswanathan, "Hands-free voice communication in an automobile with a microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1992, pp. 281–284.

[2] M. Omologo, P. Svaizer, and Matassoni, "Environmental conditions and aocustic transduction in hands-free speech recognition," *Speech Communication*, vol. 25, pp. 75–95, 1998.

[3] C. Wang, S. Griebel, P. Hsu, and M. Brandstein, "Real-time automated video and audio capture with multiple camera and microphones," *Journal of VLSI Signal Processing Systems*, vol. 29(1/2), pp. 81–100, Aug/Sep 2001.

[4] D. Zotkin, R. Duraiswami, V. Philomin, and L. Davis, "Smart videoconferencing," in *Proc. ICME2000*, August 2000, pp. 1597–1600.

[5] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 320–327, Aug. 1976.

[6] S. Bédard, B. Champagne, and A. Stéphanne, "Effects of room reverberaion on time-delay stimation performance," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1994, pp. II–261 – II–264.

[7] M. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *J. Acoust. Soc. Amer.*, vol. 105, no. 5, pp. 2914–2919, 1999.

[8] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

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

[10] 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.

[11] B. Yegnanarayana, S. R. Mahadeva Prasanna and K. Sreenivasa Rao, "Speech enhancement using excitation source information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002.

[12] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, 1981.

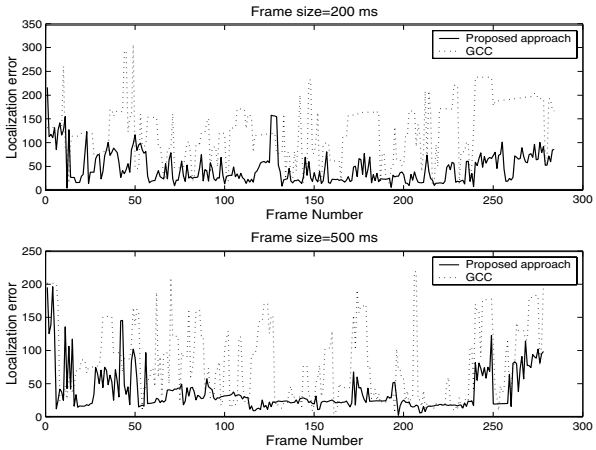

Figure 7: The Localization error as a function of frame number using the GCC approach and the proposed approach for frame length $200ms$ and $500ms$ with a frame shift of $50ms$.