Speaker Localization Using Excitation Source Information in Speech

Vikas C. Raykar, *Student Member, IEEE*, B. Yegnanarayana, *Senior Member, IEEE*, S. R. Mahadeva Prasanna, *Member, IEEE*, and Ramani Duraiswami, *Member, IEEE*

***Abstract*—This paper presents the results of simulation and real room studies for localization of a moving speaker using information about the excitation source of speech production. The first step in localization is the estimation of time-delay from speech collected by a pair of microphones. Methods for time-delay estimation generally use spectral features that correspond mostly to the shape of vocal tract during speech production. Spectral features are affected by degradations due to noise and reverberation. This paper proposes a method for localizing a speaker using features that arise from the excitation source during speech production. Experiments were conducted by simulating different noise and reverberation conditions to compare the performance of the time-delay estimation and source localization using the proposed method with the results obtained using the spectrum-based generalized cross correlation (GCC) methods. The results show that the proposed method shows lower number of discrepancies in the estimated time-delays. The bias, variance and the root mean square error (RMSE) of the proposed method is consistently equal or less than the GCC methods. The location of a moving speaker estimated using the time-delays obtained by the proposed method are closer to the actual values, than those obtained by the GCC method.**

***Index Terms*—Excitation source information, Hilbert envelope, speaker localization, time-delay estimation.**

I. Introduction and Previous Work

APPLICATIONS such as videoconferencing [1]–[3], hands-free voice communication [4], [5], speech acquisition in automobile environments [6], [7], speech recognition [8], [9], acoustic surveillance and hearing-aid devices [10] require the capture of high-quality speech from the speakers. The speech signal received from a speaker in such acoustical environments is corrupted both by additive noise and room reverberation. One effective way of dealing with such situations is to use a set of spatially distributed microphones for recording the speech. Some of the previously mentioned applications may also require localizing and tracking the moving speaker. For instance, to keep the speaker in focus in videoconferencing, the speaker can be localized, and this information can be fed to a video system for actuating the pan-tilt operations of a camera [1]–[3]. Once the actual position of the speaker is known, the microphone array can be steered electronically (beamformed) for high-quality speech acquisition. Speaker Localization is also useful in a multispeaker scenario in which speech from a particular speaker may need to be enhanced with respect to others, or with respect to noise sources.

The essential requirement for all the applications mentioned previously is the ability of the microphone array to locate a speaker accurately. Broadly three types of methods exist for localizing the speaker [11]: a) maximizing the steered response power (SRP) of a beamformer, b) methods based on high-resolution spectral estimation, and c) methods based on time difference of arrival (TDOA). In the steered beamformer approach, the microphone array is electronically steered to various locations to search for a peak in the output power. A simple delay and sum beamformer or more sophisticated beamformers which apply filtering can be used. Due to its computational complexity and lack of prior knowledge of the source and noise characteristics, this method may not be practical for localizing speakers. The second method, based on the high-resolution spectrum estimation, uses the spatio-spectral correlation matrix derived from the signals received at the microphones. The high-resolution methods are designed for far field narrow-band stationary signals and, hence, it is difficult to apply them to wide-band speech. The most commonly used method in practice is the TDOA-based method. In this method, the signals received by several microphones are processed to estimate the time-delays between pairs of microphones. The estimated time-delays can be used to derive the location of the speaker.

For effective speaker localization, it is essential to obtain a good estimate of the time-delay even when the signals are corrupted by noise and reverberation [12]. The time-delay may be estimated by locating the peak in the cross correlation function of the signals received by a pair of microphones. However, this method is not robust to degradations in the signals. Knapp and Carter [13] developed the maximum likelihood (ML) estimator for determining the time-delay between signals received at two spatially separated microphones when the noise is uncorrelated. In this method, the estimated delay is the time lag which maximizes the cross correlation between filtered versions of the received signals [13]. The cross correlation of the filtered versions

Manuscript received April 15, 2003; revised June 18, 2004. This work was supported by the National Science Foundation under Awards 0086075 and 0205271. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Peter Vary.

V. C. Raykar and R. Duraiswami are with the Perceptual Interfaces and Reality Laboratory, Institute of Advanced Computer Studies, Department of Computer Science, University of Maryland, College Park, MD 20742, USA (e-mail: vikas@umiacs.umd.edu; ramani@umiacs.umd.edu).

B. Yegnanarayana is with the Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600036, India (e-mail: yegna@cs.iitm.ernet.in).

S.R. Mahadeva Prasanna is with the Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, India (e-mail: prasanna@iitg.ernet.in).

Digital Object Identifier 10.1109/TSA.2005.851907

of the signals is called the generalized cross correlation (GCC) function. The GCC function $R_{x_1x_2}(\tau)$ is given by [13]

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

where $X_1(\omega)$ and $X_2(\omega)$ are the Fourier transforms of the microphone signals $x_1(t)$ and $x_2(t)$, respectively, and $W(\omega)$ is the weight function. The effect of five different weight functions, namely, the Roth Impulse Response, the smoothed coherence transform (SCOT), the phase transform (PHAT), the Eckart filter and the ML weighting were studied in [13].

The two most commonly used weight functions are ML and PHAT. The ML weight function accentuates the signal passed to the correlator at frequencies where the signal-to-noise ratio (SNR) is high [13]. Brandstein *et al.* [14] proposed an approximate ML type weighting for speech applications. The approximate weight function is is given by

$$\widehat{W}_{\mathrm{ML}}(\omega) = \frac{|X_1(\omega)||X_2(\omega)|}{|N_1(\omega)|^2|X_2(\omega)|^2 + |N_2(\omega)|^2|X_1(\omega)|^2} \qquad (2)$$

where $|N_1(\omega)|$ and $|N_2(\omega)|$ are the noise power spectra at the two microphones, and are assumed to be known during the silence interval [14]. We use this weight function in our simulation studies. This ML weight function performs well when the effect of room reverberation is low.

As the room reverberation increases, this method shows degradations in performance [12]. Since the spectral characteristics of the received signal are affected by the multipath propagation or reverberation in a room, the GCC function is made more robust by deemphasizing the frequency-dependent weighting. The PHAT is one extreme case where the magnitude spectrum is flattened. The PHAT weight function $W_{\mathrm{PT}}(\omega)$ is given by

$$W_{\mathrm{PT}}(\omega) = \frac{1}{|X_1(\omega)X_2^*(\omega)|}. \qquad (3)$$

By flattening the magnitude spectrum the resulting location of the peak in the GCC function corresponds to the dominant delay. However, the disadvantage of the PHAT weighting is that it places equal emphasis on both low and high SNR regions and, hence, works well only when the overall noise level is low. Stéphanne and Champagne [15] proposed cepstral prefiltering to reduce the effects of reverberation. Benesty [16] proposed a novel method for time-delay estimation based on eigenvalue decomposition of the covariance matrix.

The methods discussed previously are applicable to a general sound source. Recently, methods have been suggested for localization of speaker by modeling the production of speech [17], [18]. Brandstein [18] proposed a method based on the knowledge of the periodicity of voiced speech. This method requires the estimation of pitch and, hence, the performance depends on the robustness of pitch estimation method. Moreover, the method uses the spectral weighting based on the estimated pitch harmonics. Most of the speech-model-based methods use spectral features which correspond approximately to the characteristics of the vocal tract system during the production of speech. The spectral features are affected by transmission through medium, noise and room reverberation. Not many attempts have been made to exploit the characteristics of the excitation source during the production of speech. In this paper, we show that features based on the excitation source in speech production are robust to degradations such as noise and reverberation. We discuss methods to extract the excitation source information from a speech signal, and show how to use this information to estimate the time-delay. The proposed method does not use the periodicity property of voiced speech. The method exploits the excitation characteristics of voiced speech, especially the characteristics around the glottal closure instants.

The paper is organized as follows. A method for estimation of time-delay using the excitation source information is proposed in Section II. The proposed method is compared with GCC-PHAT, GCC-ML, and Brandstein's methods using simulations, and are discussed in Section III. In Section IV, speaker localization is described, and is compared with the results obtained using the GCC-PHAT method. The paper concludes with a summary of the present work, and a discussion on possible extensions.

II. Time-Delay Estimation Using Excitation Source Information

Speech is the result of excitation of a time-varying vocal tract system with time-varying excitation [19]. The common and significant mode of excitation of the vocal tract system is the vibration of vocal folds, called glottal vibration, which to a first approximation may be treated as consisting of a sequence of impulses [20]. The characteristics of the dynamic vocal tract system are represented by short-time spectral features. Since the signal received at a microphone is affected by noise and the response of room, the received signal contains information about the vocal tract system corrupted by different levels of degradations at different microphones. However, it is interesting to note that the relative locations of epochs or instants of significant excitation in the production of speech are not affected by degradations [21]. The epochs in a voiced segment correspond to the instants of glottal closure, and their locations along the time scale do not change with the impulse response of the acoustical environment. In unvoiced segments, also there may be epochs due to strong bursts of excitation, even though they may not occur at periodic intervals as in the voiced case. But their relative locations are unaffected by degradation.

The excitation source information can be extracted from the speech signal using linear prediction (LP) analysis [22]. In 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 speech signal sample at nth instant, then its predicted value is given by

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \qquad (4)$$

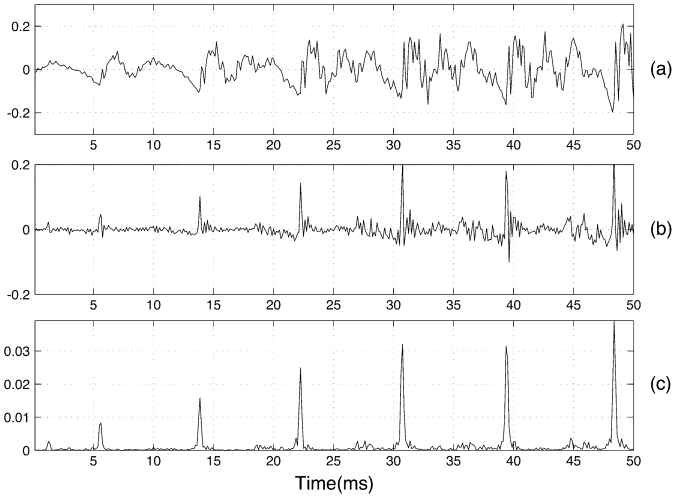

Fig. 1. (a) Speech waveform. (b) Tenth-order LP residual. (c) Hilbert envelope of the LP residual for a segment of speech signal collected over a close-speaking microphone (*mic-0*).

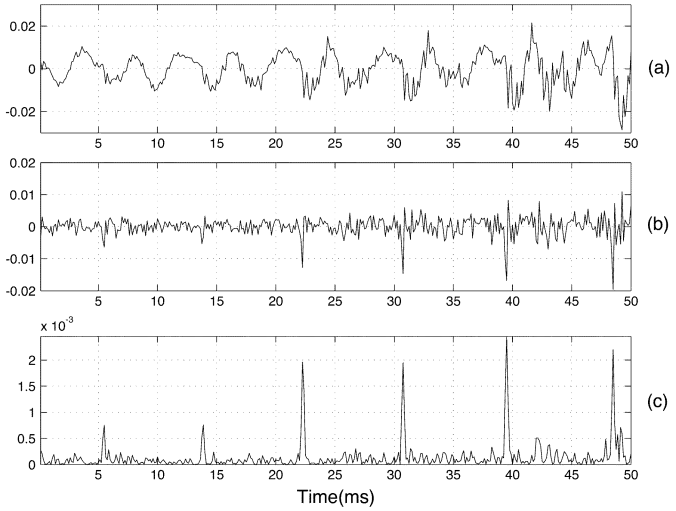

Fig. 2. (a) Speech waveform. (b) Tenth-order LP residual. (c) Hilbert envelope of the LP residual for a segment of speech signal collected over *mic-1*, which is placed at a distance of about 2.75 m from the speaker.

where $\{a_k\}$ are the LP coefficients. The error between the speech sample and its predicted value is given by

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k). \qquad (5)$$

The optimal values of the linear prediction coefficients (LPCs) can be obtained by minimizing the squared error over an analysis frame of about 10–30 ms. These LPCs define the inverse filter given by

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}. \qquad (6)$$

Passing the speech signal through this inverse filter is equivalent to using the optimal values of LPCs in (5) and, hence, the minimum error signal is the LP residual signal denoted by $e(n)$. The LP residual mostly contains information about the excitation source. The most important information about the excitation source is the sequence of epochs in the case of voiced speech.

Speech signals are collected using a microphone placed close to the speaker, which here after will be termed as *close-speaking microphone* (*mic-0*) and two other microphones (say, *mic-1* and *mic-2*), placed at a distance (*distant microphones*) in an office room of dimension $5.67 \times 4.53 \times 2.68$ m with an average reverberation time of about 0.2 s and noise level of about 40–50 dB. All the signals are sampled at 8 kHz and stored as 16 bit numbers. The microphones signals are shown in Figs. 1(a), 2(a), and 3(a), respectively. The two distant microphones are placed at a distance of about 2.75 m from the speaker. All the three signals differ from one another. The low SNR of the signals collected at the distant microphones can be seen from the amplitudes of signals in Figs. 2(a) and 3(a) in relation to the signal in Fig. 1(a). The tenth-order LP residuals derived from the speech signals of *mic-0*, *mic-1* and *mic-2* are shown in Figs. 1(b), 2(b), and 3(b), respectively. The LP residual signals in Figs. 2(b) and 3(b) also reflect the low SNR characteristics of the signals at *mic-1* and *mic-2*.

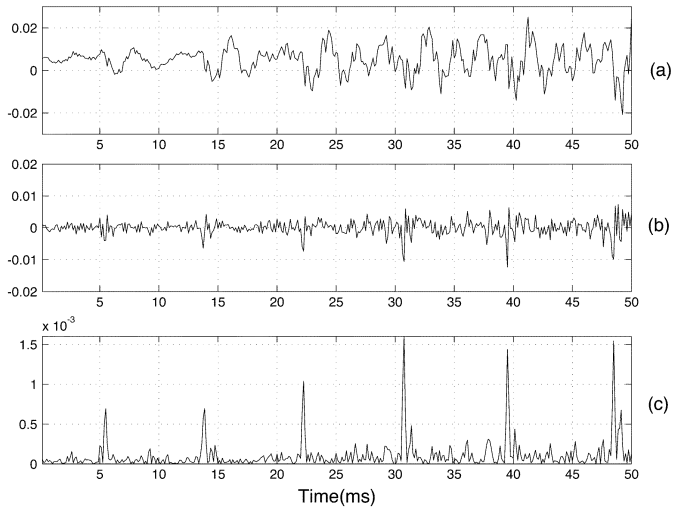

Fig. 3. (a) Speech waveform. (b) Tenth-order LP residual. (c) Hilbert envelope of the LP residual for a segment of speech signal collected over *mic-2*, which is placed at a distance of about 2.75 m from the speaker.

The time-delay may be estimated by locating the peak in the cross correlation function of signals received by two microphones. Due to degradation caused by noise and room reverberation, the signal received at one microphone will not simply be a delayed version of the other. If speech signals are directly used for computing the cross correlation function, then the correlation peak may not be prominent and distinct due to effects of noise and reverberation on the spectra of speech signals. The effects of noise and reverberation are somewhat reduced around the epochs in the LP residual, where the residual error is large. Note that the relative epoch locations are not affected by the degradations. Therefore, it is possible to obtain a peak in the cross correlation of LP residuals that corresponds mostly to the correlated components around the epochs in LP residuals. Although, due to inverse filtering, noise is enhanced in the high-frequency region in the spectrum of LP residual, this will have little effect on the peak in the cross correlation, since the noise at the two microphones are not correlated.

In each pitch period major excitation occurs at the epoch corresponding to the instant of glottal closure. Around each epoch the prediction will be poor and, hence, the error is large in the residual. However, the amplitudes of the residual signal around each epoch depend on the phase of the signal [20]. This causes random fluctuation in amplitudes, which may lead to ambiguity in the location of the peak in the cross correlation function. Therefore, instead of using the LP residual directly, the Hilbert envelope of the LP residual can be used [20]. The Hilbert envelope of the LP residual $e(n)$ is defined as

$$h(n) = \sqrt{e^2(n) + e_h^2(n)} \tag{7}$$

where $e_h(n)$ is the Hilbert transform of $e(n)$ [23]. The Hilbert transform is obtained by interchanging the real and imaginary parts of the Discrete Fourier Transform (DFT) of $e(n)$, and then taking the inverse DFT. A 1024 point DFT or higher is used throughout this study for computing the Hilbert envelope. That is, the residual signal block size is 1024 points or more for computing Hilbert envelope. Figs. 1(c), 2(c), and 3(c) show the Hilbert envelopes of the LP residuals for speech signals from *mic-0*, *mic-1* and *mic-2*, respectively. The ambiguity present around epochs in the LP residual is reduced significantly in the Hilbert envelope. The epoch locations are also clearly visible in the Hilbert envelopes of the LP residuals.

The time-delay between speech signals at a pair of microphones is estimated by computing the cross correlation of the Hilbert envelopes of the LP residuals. For every frame (size in the range 50 ms to 500 ms), the cross correlation function is computed. The choice of frame size depends on the accuracy of tracking. Smaller frame size will yield better tracking. But larger frame size will yield accurate delay estimation. In any case, each frame should contain at least a few (about 5) pitch periods to obtain good estimate of time-delay. The displacement of the peak with respect to the center of cross correlation function is the desired time-delay.

To compare different methods we define the quantity peak-to-sidelobe ratio (PSR) as the peak value divided by the standard deviation of 40 samples around the peak, excluding 5 samples on either side of the peak [24]. The PSR measure gives the strength of the main peak in relation to the values around the peak. The choice of 40 samples is quite arbitrary. Fig. 4(a) shows the cross correlation function between two 50 ms speech segments from *mic-1* and *mic-2*. The PSR values are also given in the figure. The PSR value for speech signal is 5.52. Fig. 4(b) shows the cross correlation function obtained by GCC with PHAT weighting for the same two segments [13]. It can be seen that the PSR is larger than for Fig. 4(a). The disadvantage of the PHAT weighting is that it emphasizes the noise samples and, hence, it works well only when the noise level is low. Fig. 4(c) shows the cross correlation function for the tenth-order LP residuals of the two speech segments. The plot looks similar to that for the GCC case. Fig. 4(d) shows the cross correlation function for the Hilbert envelopes of the LP residuals. The use of the Hilbert envelopes produces a significantly high value of PSR, compared to the PSR values of the three previous cases. This is because, in the Hilbert envelopes of the LP residuals, the high SNR portions correspond to the major excitations

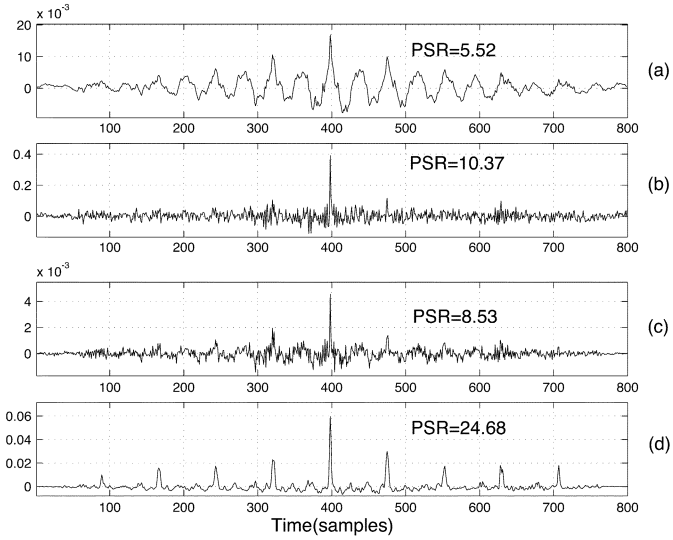

Fig. 4. Cross correlation function for different cases for 50 ms voiced speech segments from *mic-1* and *mic-2*. (a) Speech signals. (b) GCC with PHAT weighting. (c) Tenth-order LP residuals. (d) Hilbert envelopes of the LP residuals. PSR is computed for the largest peak in each cross correlation function.

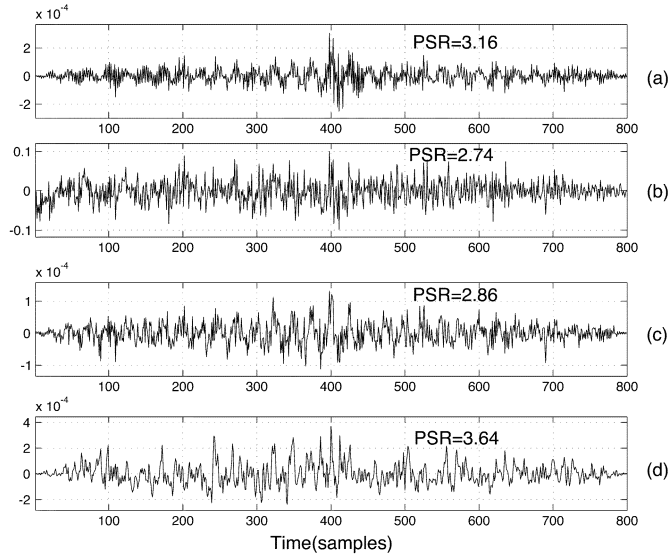

Fig. 5. Cross correlation function for different cases for 50 ms unvoiced speech segments from *mic-1* and *mic-2*. (a) Speech signals. (b) GCC with PHAT weighting. (c) Tenth-order LP residuals. (d) Hilbert envelopes of the LP residuals. PSR is computed for the largest peak in each cross correlation function.

(epochs) of the vocal tract system. The high-amplitude values at the epochs in the signal dominate the computation of the cross correlation function. Note that the time-delay is estimated using only the main peak in the cross correlation function. The other large peaks in Fig. 4(d) are due to the pitch period. Since the PSR value computed from the Hilbert envelopes of the LP residuals is high for a given voiced segment, we use the PSR value for each frame to derive a normalized weight function in order to compare the bias, variance and root mean square error (RMSE) for each of the methods.

Fig. 5 shows the cross correlation functions for a 50 ms unvoiced segment. Even for unvoiced segment the PSR value is

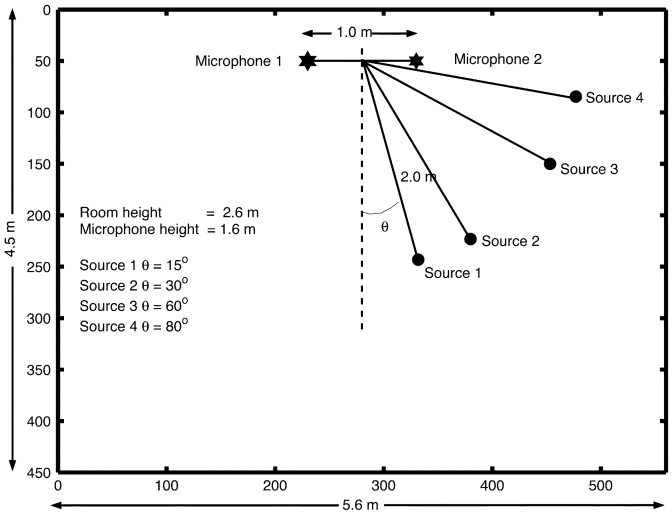

Fig. 6. Top view of the simulated room used to evaluate the proposed time-delay estimation method.

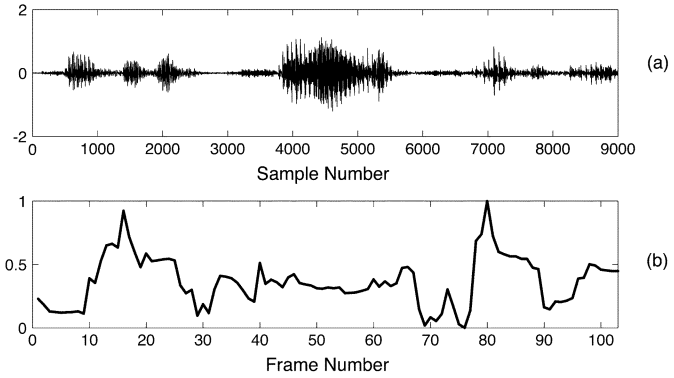

Fig. 7. (a) Sample speech waveform with reverberation of 100 ms and SNR 30 dB. (b) Corresponding PSR weighting function for a framesize of 100 ms with a shift of 10 ms. The PSR was computed for the proposed method.

high when Hilbert envelope of LP residual is used. But the PSR value depends on the strength of the bursts in the unvoiced segment. Note that the bursts need not be periodic. Hence, for unvoiced segments also the Hilbert envelope is useful for obtaining a correlated peak with PSR value higher than other methods.

III. Comparison With Other Methods

In this section, time-delays estimated using the excitation source information are compared with those obtained from other methods. In particular, we compare the results by the proposed method with the results from the GCC with PHAT weighting [13], GCC with ML weighting [14], and Brandstein's pitch-based weighting [18] methods. The relative performance of the proposed method is evaluated using a series of Monte Carlo trials in a simulated rectangular room of dimension $5.6 \times 4.5 \times 2.6$ m as illustrated in Fig. 6. The microphones are assumed to have an omnidirectional pattern. The source is placed at a distance of 2.0 m from the center of microphone pair which are 1 m apart. Simulation studies are made for four different source positions, each corresponding to a different direction of arrival (DOA) as shown in Fig. 6. The DOA is the angle between the line joining the source to the center of the microphone pair, and the normal to the line joining the two microphones at the center of the microphone pair. The four positions of the source shown in Fig. 6 correspond to DOAs of 15°, 30°, 60°, and 80°. The simulated walls are plane reflective surfaces with frequency independent reflection coefficients. The impulse response between any two points in the room is generated using Allen and Berkley's image method [25]. The impulse response is convolved with the input signal to simulate the effect of room reverberation.[1] The simulation studies are carried out for reverberation times varying from 0 to 0.3 s. The reflection coefficient β for a given room dimension and reverberation time are related by the Eyring's formula $\beta = \exp(-13.82/c[L_x^{-1} + L_y^{-1} + L_z^{-1}]T)$ [26], where $L_x, L_y,$ and L_z are the dimensions of the room, T is reverberation time in seconds and c is speed of sound in air (342 m/s).

Speech recorded over a close-speaking microphone in noise free conditions and sampled at 8 kHz is used in these studies. The speech signal is convolved with the impulse response of the room to derive the reverberant signal. The SNR of reverberant signal is then varied from 0 to 50 dB by adding zero mean white Gaussian noise to the speech signal. The resulting degraded speech signal is segmented into frames of 200 ms with a shift of 50 ms. Each segment is multiplied with a Hanning window [19]. The time-delay is estimated for each frame using the proposed method, and by the GCC method with PHAT, ML, and Brandstein's pitch-based[2] weighting.

The performance of the time-delay estimation method is evaluated by calculating the bias, variance and RMSE for different room impulse responses and SNR values. In each of the simulations, the actual time-delay can be calculated corresponding to a given DOA. Often noise and some unvoiced segments give large random error, and thus these segments contribute significantly to the estimated bias, variance and RMSE. To reduce the contribution due to these segments, the knowledge of the PSR value of each frame is used. The PSR values are relatively high in voiced regions, and low in some unvoiced and noise regions. The PSR values computed by the proposed method are used for deriving a weight function. A sample weight function is shown in Fig. 7(b) for the speech waveform shown in Fig. 7(a). The errors in the estimated time-delays by all the four methods are weighted for computing the bias, variance and RMSE values. The bias, variance and RMSE values given for different cases are computed by averaging the results obtained from 100 different simulations.

Figs. 8–10 show the bias (in number of samples), variance (in number of samples square) and RMSE (in number of samples), respectively, for a DOA of 15°. The SNR and the reverberation time, respectively, are varied from 0 to 50 dB and 0 to 0.3 s. For very low SNR, the GCC-ML performs better than all the other methods (see 0–10 dB regions in all the plots). The GCC-ML weighting has been derived as the optimal estimator when the noise is Gaussian. Since in our simulations we use the Gaussian

[1]The nonphysical behavior of the Allen and Berkley's image method at zero frequency is avoided by using a low cutoff (1% of the sampling frequency) high-pass filter [25].

[2]For the Brandstein's pitch-based method [18] we estimate the pitch directly from the clean speech signal rather than the reverberant noisy signal. As a result there will not be errors due to error in the pitch estimation.

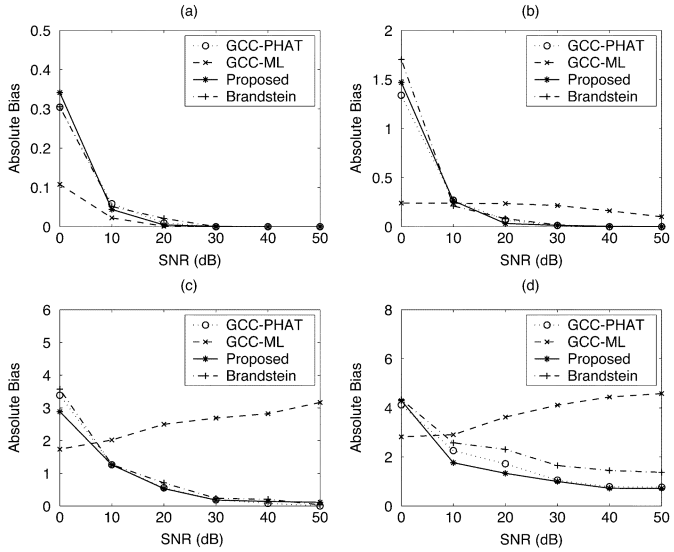

Fig. 8. Comparison of absolute bias (in number of samples) for the four methods: GCC-PHAT, GCC-ML, Brandstein's pitch-based, and the proposed method. The DOA is 15° and the SNR is varied from 0 dB to 50 dB. Four different reverberation times are considered. (a) Reverberation time = 0.0 s. (b) Reverberation time = 0.1 s. (c) Reverberation time = 0.2 s. (d) Reverberation time = 0.3 s. The scale on y axis in each of the subplots is different.

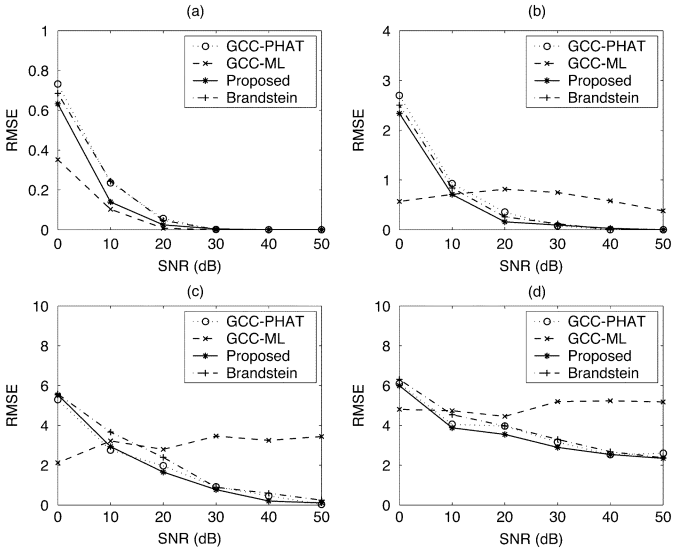

Fig. 10. Comparison of RMSE (in number of samples) for the four methods: GCC-PHAT, GCC-ML, Brandstein's pitch-based, and the proposed method. The DOA is 15° and the SNR is varied from from 0 dB to 50 dB. Four different reverberation times are considered. (a) Reverberation time = 0.0 s. (b) Reverberation time = 0.1 s. (c) Reverberation time = 0.2 s. (d) Reverberation time = 0.3 s. Note that the scale on y axis in each of the subplots is different.

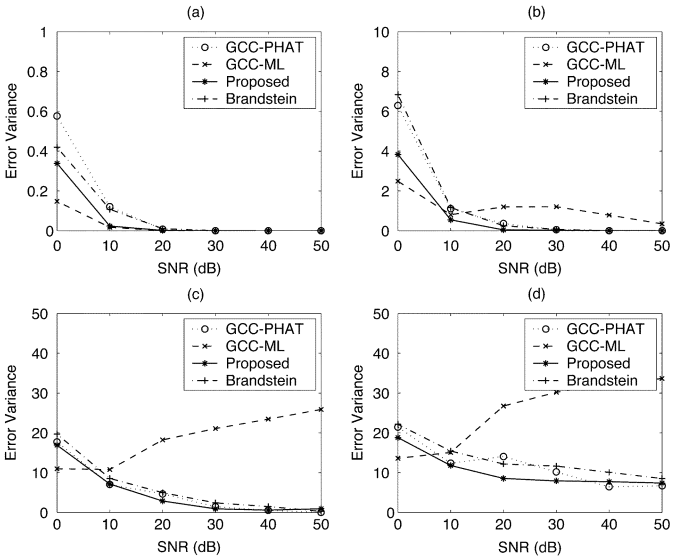

Fig. 9. Comparison of the error variance (in number of samples square) for the four methods: GCC-PHAT, GCC-ML, Brandstein's pitch based, and the proposed method. The DOA is 15° and the SNR is varied from 0 dB to 50 dB. Four different reverberation times are considered. (a) Reverberation time = 0.0 s. (b) Reverberation time = 0.1 s. (c) Reverberation time = 0.2 s. (d) Reverberation time = 0.3 s. The scale on y axis in each of the subplots is different.

noise model, it is not surprising that GCC-ML performs the best. For high SNR and low reverberation, GCC-ML, GCC-PHAT and the Brandstein's pitch-based method perform equally well. The Brandstein's pitch-based method performs slightly better than the GCC-PHAT method, and the GCC-PHAT performs better than the GCC-ML. The proposed method performs better than all these three methods [see, 20–50 dB regions in Figs. 8, 9, and 10(a), (b)]. For low SNR and high reverberation GCC-ML seems to be performing better than GCC-PHAT (see, 0–10 dB regions in Figs. 8, 9, and 10(c), (d)]. For high SNR and high reverberation the proposed method outperforms all the other three methods (see, 10–50 dB regions in Figs. 8, 9, and 10(c), (d)]. Thus, it can be concluded that the performance of the proposed method is consistently equal to, or better than, the best performing of the three methods.

One more metric, namely, *percentage discrepancy* is introduced, which is defined as the percentage of trials for which the absolute error in the estimated delay is greater than a given threshold ($\pm 20°$ in the DOA). Fig. 11 shows percentage discrepancies in the estimated delays for the proposed and the GCC methods for the DOA corresponding to 15°. From Fig. 11(a), it can be seen that all the three methods perform equally well for the zero reverberation case. As the reverberation increases, the GCC-PHAT method gives lower discrepancies compared to the GCC-ML method for high SNR values. The proposed method gives significantly fewer discrepancies for all the SNR values.

Similar trends in bias, variance, RMSE and percentage discrepancies were observed for the experiments with DOAs 30° and 60°. For illustration, we have given the RMSE for the case of reverberation time of 0.3 s in Fig. 12. Similar experiments were conducted using colored noise obtained by bandpass filtering the white noise. Figs. 13 and 14 show the RMSE and percentage discrepancies, respectively, for a DOA of 15° for colored noise. In all these cases, the proposed method performs better than other methods. For the bandpass filtered noise the GCC-ML performs consistently worse than the other methods.

IV. Localization of Speaker in a Real Environment

Localization of speaker in an acoustical environment involves two steps. The first step is estimation of time-delays between pairs of microphones. The next step is to use these delays to estimate the location of speaker.

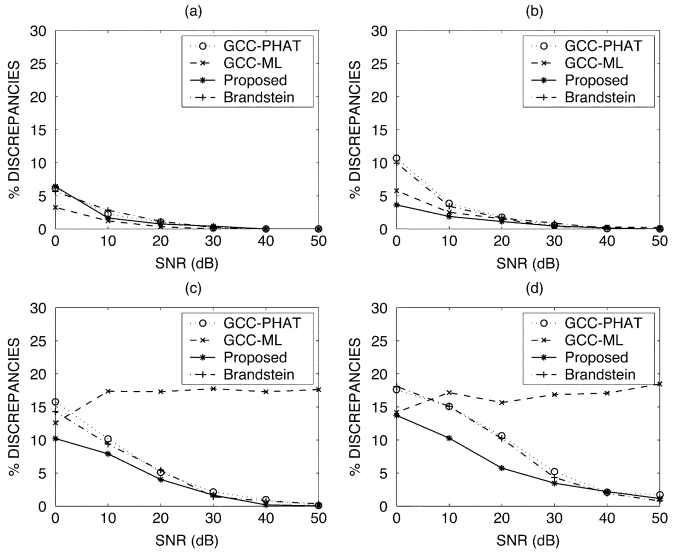

Fig. 11. Comparison of percentage discrepancies for the four methods: GCC-PHAT, GCC-ML, Brandstein's pitch-based, and the proposed method. The DOA is 15° and the SNR is varied from 0 dB to 50 dB. Four different reverberation times are considered. (a) Reverberation time = 0.0 s. (b) Reverberation time = 0.1 s. (c) Reverberation time = 0.2 s. (d) Reverberation time = 0.3 s.

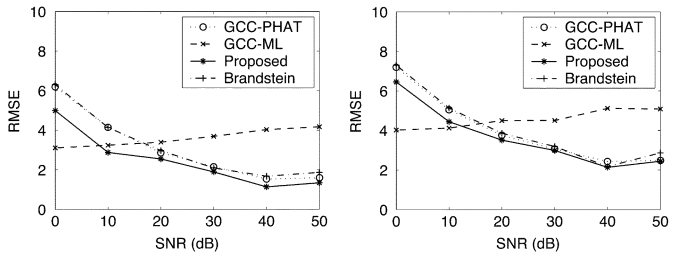

Fig. 12. Comparison of RMSE (in number of samples) for the four methods: GCC-PHAT, GCC-ML, Brandstein's pitch-based, and the proposed method for reverberation time 0.3 s corresponding to the DOA. (a) 30°. (b) 60°.

The speaker localization problem may 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 spatial coordinates (x, y and z) of two microphones in the ith pair. Let the source be located at $\mathbf{s}$. The actual delay associated with a source at $\mathbf{s}$ and the ith pair of microphones is given by

$$t_i(\mathbf{s}) = \frac{|\mathbf{s} - \mathbf{m_i^1}| - |\mathbf{s} - \mathbf{m_i^2}|}{c} \tag{8}$$

where c is the speed of propagation of sound ($c = 342$ ms^{-1} at room temperature). The speed of sound in a given acoustical medium is assumed to be constant. Let τ_i be the estimated time-delay. If the estimated time-delay is corrupted by zero-mean additive white Gaussian noise with known variance $v(\tau_i)$, then τ_i is normally distributed with mean $t_i(\mathbf{s})$ and variance $v(\tau_i)$

$$\tau_i \sim \mathbf{N}(t_i(\mathbf{s}), v(\tau_i)). \tag{9}$$

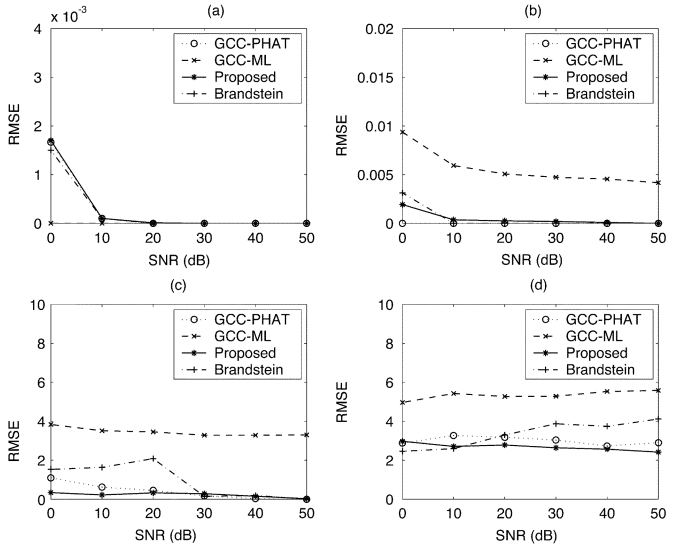

Fig. 13. Comparison of RMSE (in number of samples) for the four methods: GCC-PHAT, GCC-ML, Brandstein's pitch-based, and the proposed method. The DOA is 15° and the SNR is varied from from 0 dB to 50 dB. Four different reverberation times are considered. (a) Reverberation time = 0.0 s. (b) Reverberation time = 0.1 s. (c) Reverberation time = 0.2 s, and (d) Reverberation time = 0.3 s. Colored noise was used for these results. Note that the scale on y axis in each of the subplots is different.

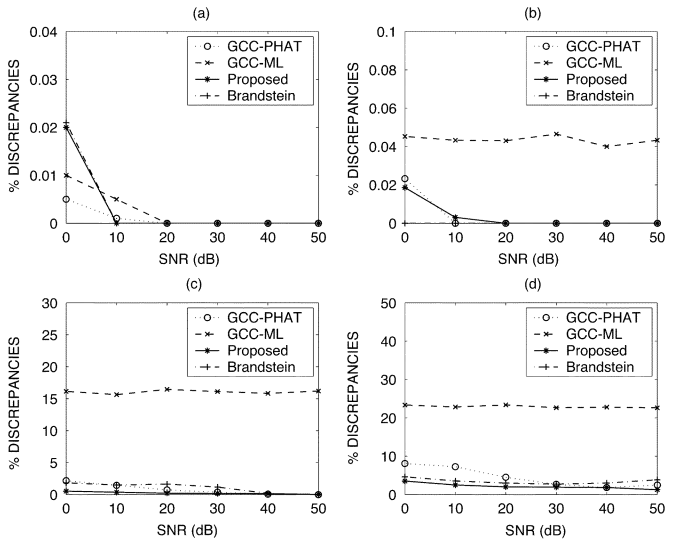

Fig. 14. Comparison of percentage discrepancies for the four methods: GCC-PHAT, GCC-ML, Brandstein's pitch-based, and the proposed method. The DOA is 15° and the SNR is varied from 0 dB to 50 dB. Four different reverberation times are considered. (a) Reverberation time = 0.0 s. (b) Reverberation time = 0.1 s. (c) Reverberation time = 0.2 s. (d) Reverberation time = 0.3 s. Colored noise was used for these results. Note that the scale on y axis in each of the subplots is different.

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

$$p(\tau_1, \tau_1, \ldots, \tau_M; \mathbf{s}) = \prod_{i=1}^{M} \frac{1}{\sqrt{2\pi v(\tau_i)}} \exp\left[\frac{-(\tau_i - t_i(\mathbf{s}))^2}{2v(\tau_i)}\right]. \tag{10}$$

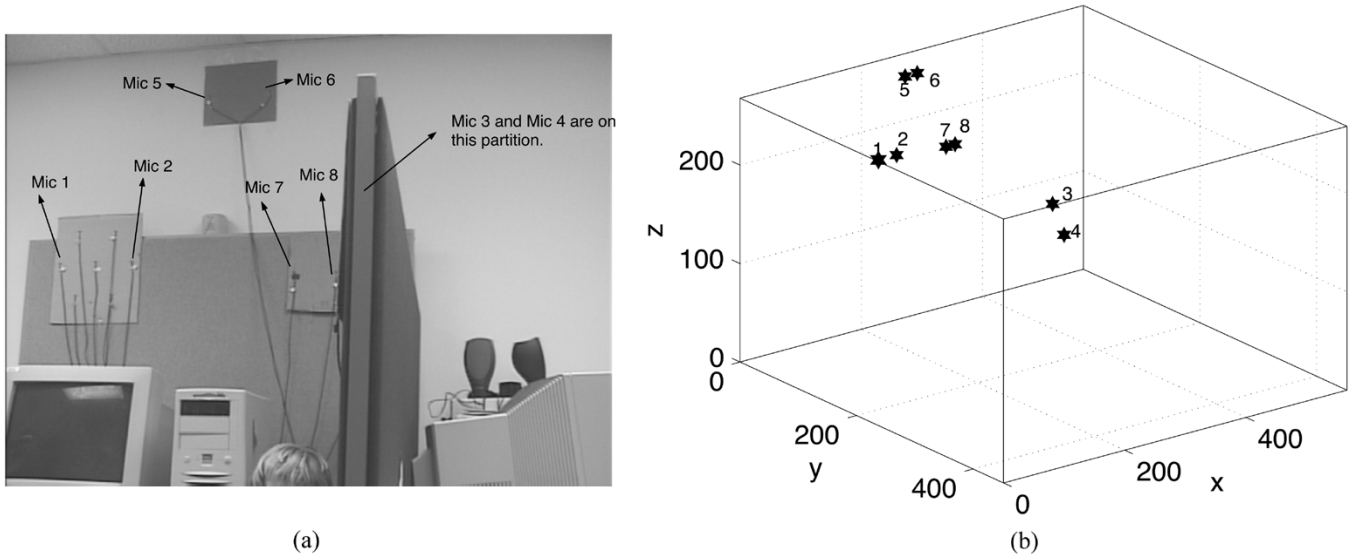

Fig. 15. (a) Picture showing the actual setup of the microphones. (b) Schematic of the room indicating the positions of the eight microphones selected for the study. Microphones 3 and 4 are on the partition.

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

$$J_{\mathrm{ML}}(\mathbf{s}) = \sum_{i=1}^{M} \frac{(\tau_i - t_i(\mathbf{s}))^2}{v(\tau_i)}. \tag{11}$$

This does not have a closed-form solution for the source position, since it is a nonlinear function of $\mathbf{s}$. Nonlinear optimization methods are needed to solve this problem. In our experiments, we used the Gauss–Newton nonlinear least square method to minimize this function [27]. The initial guess was set at the center of the room.

In order to study the effectiveness of the proposed method for speaker localization in noisy and reverberant environment, an 8 element microphone array is setup in an office room of dimension $5.67 \times 4.53 \times 2.68$ m. The reverberation time of the room is approximately 0.2 s, and the noise level in the room was about 40–50 dB. Fig. 15(a) shows the actual microphone setup in the room, and Fig. 15(b) shows the schematic of room and the positions of microphones.[3]

For all the experiments speaker was instructed to move in the room reading a text at his normal level of speaking. In order to validate the results, speaker was asked to move in a predetermined path with known coordinates. The actual path for his movement was marked on the floor of room. The speaker moved in such a way that he was always facing the microphones. In each case, as the speaker moved, the *localization error*, defined as the distance between the actual position of speaker and the estimated position of speaker, was plotted. The delays were estimated using the proposed method and the GCC-PHAT method. Frame lengths of 200 ms and 500 ms, each with a shift of 50 ms were used.

[3]The microphones are electret microphones. Data acquisition is done using the Power DAQ board PD-MF-16-330/12L. The microphones are connected to the board through a custom-built preamplifier. Signal from each channel is sampled at 8 kHz sampling frequency.

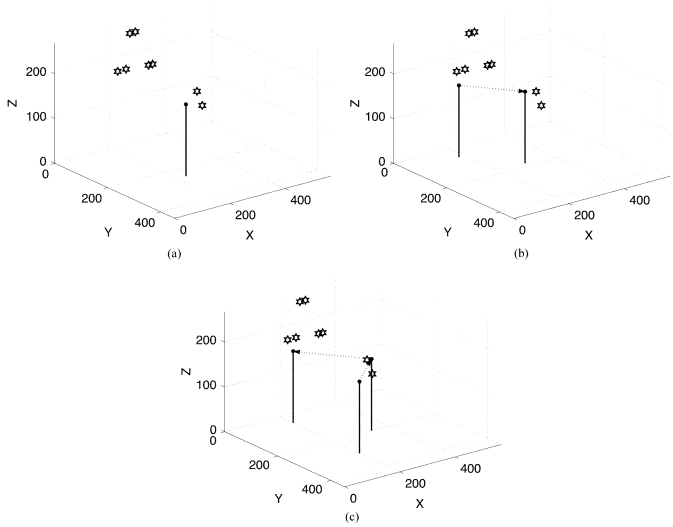

Fig. 16. Three cases for which the methods were tested. (a) Case 1: Speaker is stationary. (b) Case 2: Speaker moves from one end of the room toward the microphones. (c) Case 3: Speaker moves from one end of the room toward the microphones and from the microphones toward the other end of the room.

The following three cases were considered for study: 1) Stationary speaker. 2) Speaker moving from one end of room toward the microphones. 3) Speaker moving from one end of room toward the microphones, and then from the microphones toward the other end of room. Fig. 16 shows all three cases.

Fig. 17 shows the estimated delays as a function of frame number for one microphone pair (*mic-1* and *mic-4*) for Case 2 in Fig. 16, using the proposed and GCC-PHAT methods for frame lengths of 200 ms and 500 ms, with a frame shift of 50 ms. It can be seen that delay values vary in accordance with the movement of speaker, though there are a few random delays. Also it can be seen that the number of random delays are reduced as the frame size is increased, giving a better estimate of delays. In particular, the number of random delays obtained using the proposed method are less as compared to the GCC-PHAT method.

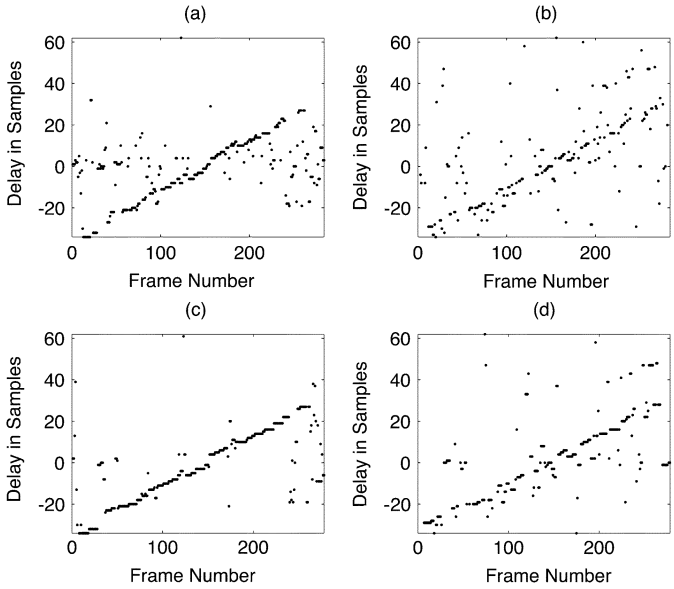

Fig. 17. Estimated delay as a function of frame number for one microphone pair (*mic-1* and *mic-4*) using (a) proposed and (b) GCC methods using a frame size of 200 ms, and (c) proposed and (d) GCC methods using a frame size of 500 ms, both with a frame shift of 50 ms for Case 2 [shown in Fig. 16(b)].

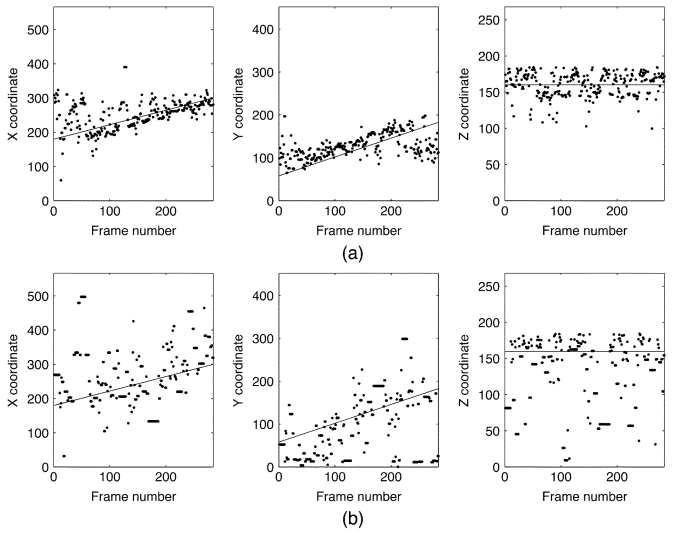

Fig. 18. Actual and the estimated locations (x, y, and z coordinates in cm) of the speaker. (a) Proposed method. (b) GCC method. A frame size of 200 ms and a frame shift of 50 ms were used for the Case 2 shown in Fig. 16(b). The actual path is shown as solid line, and the estimated path is shown as dots.

Figs. 18 and 19 show the actual and the estimated (x, y, z) coordinates for Case 2 by proposed and GCC-PHAT methods, for frame lengths of 200 ms and 500 ms, respectively. In these plots, the actual path is shown using a solid line, and the estimated path is shown using dots. It can be seen that the estimated path follows the actual path more closely for the proposed method than for the GCC-PHAT method. Figs. 20 and 21 show the localization error as a function of frame number using the proposed and GCC-PHAT methods for Case 2 and Case 3. From these plots, it can be observed that, for a given frame size, the localization error is lower for the proposed method compared to the error obtained by the GCC-PHAT method. The error is generally lower

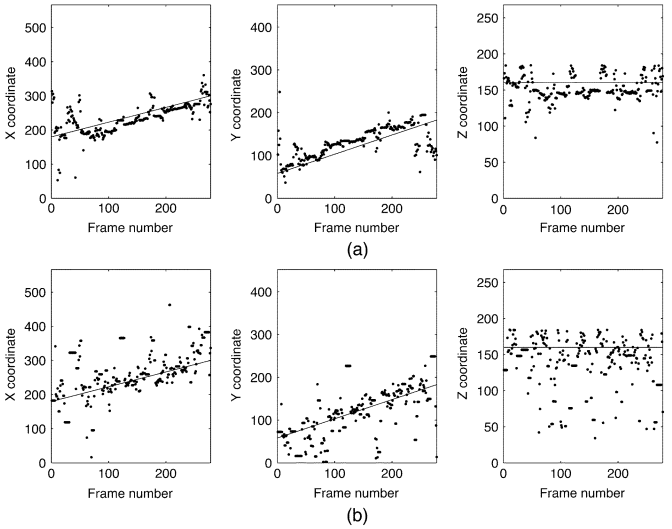

Fig. 19. Actual and the estimated location (x, y, and z coordinates in cm) of the speaker. (a) Proposed method. (b) GCC method. A frame size of 500 ms with frame shift of 50 ms were considered for the Case 2 as shown in Fig. 16(b). The actual path is shown as solid line and the estimated path is shown as dots.

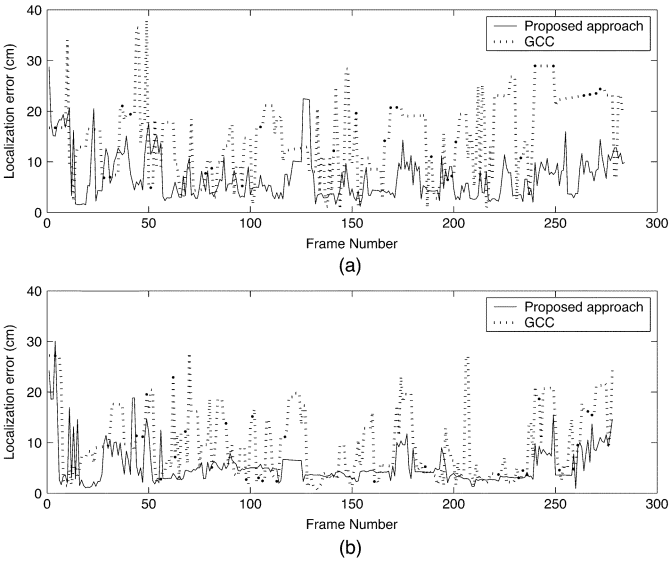

Fig. 20. Localization error (in cm) as a function of frame number using the proposed and GCC methods for frame size of (a) Frame size $=$ 200 ms and (b) Frame size $=$ 500 ms with frame shift of 50 ms for the Case 2 as shown in Fig. 16(b).

for frames where signal energy is high, and also a lower error is obtained when larger frame sizes are used.

V. Conclusion

In this paper, a method for estimation of time-delays and speaker localization using the information in the excitation source of speech production was proposed. Comparison of the results show that the delay and location parameters estimated by the proposed method are closer to the actual values than the parameters estimated from the spectral-based GCC method. Generally all the correlation-based methods work better when longer segments are used. The proposed method works even with smaller segments. Since the proposed method is based on the information in the source of excitation, the Hilbert envelope

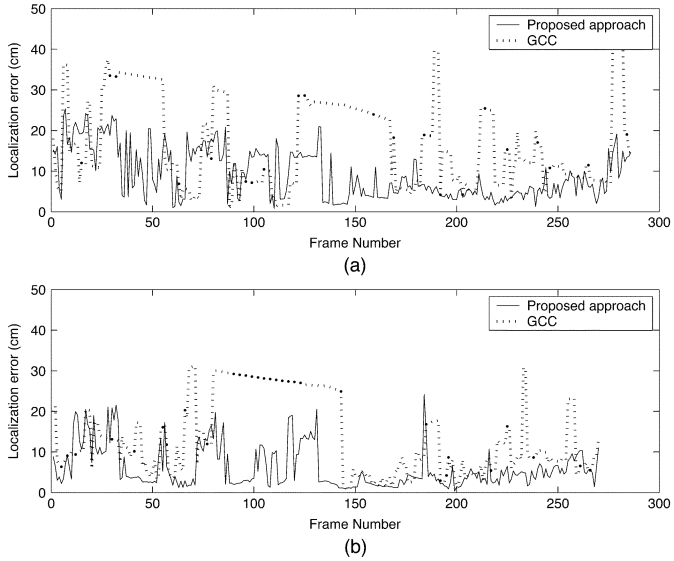

Fig. 21. Localization error (in cm) as a function of frame number using the proposed and GCC methods for frame size of (a) frame size = 200 ms and (b) frame size = 500 ms with frame shift of 50 ms for the Case 3 as shown in Fig. 16(c).

of the LP residual of even four or five pitch periods may be sufficient for estimating time-delays. In general, features of the vocal tract system and features of the excitation source contain significant information about a moving speaker. The potential of vocal tract system features has already been established. In this paper the usefulness of excitation source information is illustrated. An effective way of combining these two approaches may yield a robust method for localization and tracking a moving speaker in an adverse acoustic environment.

REFERENCES

[1] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Orlando, FL, Apr. 1997, pp. 187–190.

[2] C. Wang, S. Griebel, P. Hsu, and M. Brandstein, "Real-time automated video and audio capture with multiple camera and microphones," *J. VLSI Signal Process. Syst.*, vol. 29, no. 1/2, pp. 81–100, Aug./Sep. 2001.

[3] D. Zotkin, R. Duraiswami, V. Philomin, and L. Davis, "Smart videoconferencing," in *Proc. Int. Conf. Multimedia Expo.*, New York, Aug. 2000, pp. 1597–1600.

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

[5] R. L. B. Jeannès, P. Scalart, G. Faucon, and C. Beaugeant, "Combined noise and echo reduction in hands-free systems: A survey," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 808–820, Nov. 2001.

[6] N. Kawaguchi, S. Matsubara, H. Iwa, S. Kajita, K. Takeda, F. Itakura, and Y. Inagaki, "Construction of speech corpus in moving car environment," in *Proc. Int. Conf. Spoken Language Processing*, vol. III, Beijing, China, 2000, pp. 362–365.

[7] S. Nordholm, I. Claesson, and N. Grbiæ, *Optimal and Adaptive Microphone Arrays for Speech Input in Automobiles*, ser. Microphone arrays-Signal Processing Techniques and Applications. Berlin, Germany: Springer-Verlag, 2001, ch. 14, pp. 307–329.

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

[9] M. Omologo, M. Matassoni, and P. Svaizer, *Speech Recognition with Microphone Arrays*, ser. Microphone arrays-Signal Processing Techniques and Applications. Berlin, Germany: Springer-Verlag, 2001, ch. 15, pp. 331–353.

[10] J. E. Greenberg and P. M. Zurek, *Microphone-Array Hearing Aids*, ser. Microphone arrays-Signal Processing Techniques and Applications. Berlin, Germany: Springer-Verlag, ch. 11, pp. 229–253.

[11] J. DiBiase, H. Silverman, and M. Brandstein, *Robust Localization in Reverberant Rooms*, ser. Microphone arrays-Signal Processing Techniques and Applications. Berlin, Germany: Springer-Verlag, 2001, ch. 8, pp. 157–180.

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

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

[14] M. Brandstein, J. Adcock, and H. Silverman, "A practical time-delay estimator for localizing speech sources with a microphone array," *Comput. Speech Lang.*, vol. 9, pp. 153–169, Sep. 1995.

[15] A. Stéphanne and B. Champagne, "Cepstral prefiltering for time delay estimation in reverberant environments," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 5, Detroit, MI, May 1995, pp. 3055–3058.

[16] J. Benesty, "Adaptive eigen value decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 384–391, 2000.

[17] M. Brandstein and S. Griebel, *Explicit Speech Modeling for Microphone Arrays*, ser. Microphone arrays-Signal Processing Techniques and Applications. Berlin, Germany: Springer-Verlag, 2001, ch. 7, pp. 133–153.

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

[19] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[20] T. V. 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, no. 4, pp. 309–319, Aug. 1979.

[21] P. S. Murthy and B. Yegnanarayana, "Robustness of group-delay-based method for extraction of significant instants of excitation from speech signals," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 6, pp. 609–619, Nov. 1999.

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

[23] B. Yegnanarayana, S. R. M. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. I, Orlando, FL, May 2002, pp. 541–544.

[24] B. V. K. V. Kumar, M. Savvides, K. Venkataramani, and C. Xie, "Spatial frequency domain image processing for biometric recognition," in *Proc. IEEE Int. Conf. Image Processing*, vol. I, Rochester, NY, 2002, pp. 53–56.

[25] J. B. Allen and D. Berkely, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.

[26] C. F. Eyring, "Reverberation time in dead rooms," *J. Acoust. Soc. Amer.*, vol. 1, pp. 217–241, 1930.

[27] G. J. Borse, *Numerical Methods with MATLAB*. Boston, MA: PWS, 1998.

Vikas C. Raykar (S'00) received the B.S. degree in electronics and communication engineering from the National Institute of Technology (formerly Regional Engineering College), Trichy, India, in 2001, the M.S. degree in electrical engineering from the University of Maryland, College Park, in 2003, and is currently pursuing the Ph.D. degree in computer science from the same university.

He is currently working as a Research Assistant at the Perceptual Interfaces and Reality Laboratory, Institute of Advanced Computer Studies, University of Maryland, College Park. His research interests broadly span computer audition/vision and specifically include spatial audio, auditory source localization and microphone/camera array calibration.

B. Yegnanarayana (M'78–SM'84) was born in India in 1944. He received the B.E., M.E., and Ph.D. degrees in electrical communication engineering from the Indian Institute of Science, Bangalore, India, in 1964, 1966, and 1974, respectively.

He was a Lecturer from 1966 to 1974 and an Assistant Professor from 1974 to 1978, in the Department of Electrical Communication Engineering, Indian Institute of Science. From 1978 to 1980, he was a Visiting Associate Professor of computer science at Carnegie Mellon University, Pittsburgh, PA. Since 1980, he has been a Professor in the Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, India. He was the Chairman of the department from 1985 to 1989. His current research interests are in signal processing, speech, vision, neural networks, and man-machine interfaces. He has published papers in reviewed journals in these areas.

Dr. Yegnanarayana is a Fellow of Indian National Science Academy, Indian National Academy of Engineering, and Indian Academy of Sciences. He is an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING.

S. R. Mahadeva Prasanna (M'05) was born in India in 1971. He received the B.E. degree in electronics engineering from Sri Siddartha Institute of Technology, Bangalore University, India, in 1994, the M.Tech. degree in industrial electronics from National Institute of Technology, Surathkal, India, in 1997, and the Ph.D. degree in computer science and engineering from the Indian Institute of Technology Madras, Chennai, in 2004.

He is currently an Assistant Professor in the Department of of Electronics and Communication Engineering, Indian Institute of Technology, Guwahati. His research interests are in speech signal processing and neural networks.

Ramani Duraiswami (M'99) received the B.Tech. degree in mechanical engineering from the Indian Institute of Technology, Bombay, India, in 1985 and the Ph.D. degree in mechanical engineering and applied mathematics from The Johns Hopkins University, Baltimore, MD, in 1991.

He is currently a Faculty Member in the Department of Computer Science, Institute for Advanced Computer Studies (UMIACS), University of Maryland, College Park. He is the Director of the Perceptual Interfaces and Reality Laboratory there. His research interests are broad and currently include spatial audio, virtual environments, microphone arrays, computer vision, statistical machine learning, fast multipole methods, and integral equations.