ADAPTIVE EIGENVALUE DECOMPOSITION ALGORITHM FOR REALTIME ACOUSTIC SOURCE LOCALIZATION SYSTEM

Yiteng Huang

329506 Georgia Tech Station
Atlanta, Georgia 30332
gt9506b@prism.gatech.edu

Jacob Benesty, Gary W. Elko

Bell Laboratories, Lucent Technologies
700 Mountain Avenue
Murray Hill, New Jersey 07974
jbenesty@bell-labs.com
gwe@.bell-labs.com

ABSTRACT

To locate an acoustic source in a room, the relative delay between microphone pairs must be determined efficiently and accurately. However, most traditional time delay estimation (TDE) algorithms fail in reverberant environments. In this paper, a new approach is proposed that takes into account the reverberation of the room. A realtime PC-based TDE system running under Microsoft$^{\mathrm{TM}}$ Windows system was developed with three TDE techniques: classical cross-correlation, phase transform, and a new algorithm that is proposed in this paper. The system provides an interactive platform that allows users to compare performance of these algorithms.

1. INTRODUCTION

Realtime acoustic source localization system can be used in such applications as camera pointing for teleconferencing and microphone array beamformer steering for audio communication and speech processing systems. The problem is difficult because of the nonstationarity of speech and of room acoustic reverberation. Over the last two decades, several approaches have been proposed. Time delay estimation (TDE) between two microphones is becoming the technique of choice, especially in recent digital systems.

Generalized cross-correlation (GCC) [1] is the most commonly used method for TDE. In this technique, the delay estimate is obtained as the time-lag that maximizes the cross-correlation between filtered versions of the received signals. Techniques have been proposed to improve the GCC in the presence of noise [2, 3]. Because GCC is based on an ideal signal propagation model, it is believed that it has a fundamental weakness of inability to cope well in reverberant environments as shown clearly in [4]. Some improvement may be gained by cepstral prefiltering [5], however, shortcomings still remain. Even though more sophisticated techniques [6] exist, they tend to be computationally intensive and are thus not well suited for real-time applications.

In this paper, a new approach is proposed that is based on a real signal propagation model (with reverberation) using eigenvalue decomposition. Indeed, it will be shown that the eigenvector corresponding to the minimum eigenvalue of the covariance matrix of the microphone signals contains the impulse responses between the source and the microphones (and therefore all the information we need for TDE).

In order to evaluate consistent and dynamic performance of proposed algorithm over a range of representative acoustic conditions, a real-time acoustic source localization system was developed running on the Windows$^{\mathrm{TM}}$ 95/NT operating systems. Three methods were implemented, namely classical cross-correlation, phase transform, and the proposed adaptive eigenvalue decomposition algorithm.

2. MODELS FOR THE TDE PROBLEM

2.1. Ideal Free-Field Model

For the given source signal $s(n)$ propagating through a generic noisy free space, the signal acquired by the i-th ($i = 1, 2$) microphone can be expressed as follows:

$$x_i(n) = \alpha_i s(n - \tau_i) + b_i(n), \tag{1}$$

where α_i is an attenuation factor due to propagation loss, τ_i is the propagation time and $b_i(n)$ is the additive noise. It is further assumed that $s(n)$, $b_1(n)$, and $b_2(n)$ are zero-mean, uncorrelated, stationary Gaussian random processes. The relative delay between the two microphone signals 1 and 2 is defined as

$$\tau_{12} = \tau_1 - \tau_2. \tag{2}$$

This model generates mathematically clear solution for τ_{12} and is widely used for the classical TDE problem.

2.2. Real Reverberant Model

Unfortunately, in a real acoustic environment we must take into account the reverberation of the room and the ideal model no longer holds. Then, a more complicated but more complete model for the microphone signals $x_i(n), i = 1, 2$ can be expressed as follows:

$$x_i(n) = g_i * s(n) + b_i(n), \tag{3}$$

where $*$ denotes convolution and g_i is the acoustic impulse response of the channel between the source and the i-th microphone. Moreover, $b_1(n)$ and $b_2(n)$ might be correlated which is the case when the noise is directional, e.g., from a ceiling fan or an overhead projector.

In this case, we do not have an "ideal" solution to the problem, as is the case for the previous model, unless we can accurately determine the two impulse responses, which is a very challenging problem.

3. THE GCC METHOD

In the GCC technique, which is based on the ideal signal propagation model, the time-delay estimate is obtained as the value of τ that maximizes the generalized cross-correlation function given by

$$\begin{aligned}\psi_{x_1x_2}(\tau) &= \int_{-\infty}^{+\infty} \Phi(f) S_{x_1x_2}(f) \mathrm{e}^{j2\pi f\tau} df \\ &= \int_{-\infty}^{+\infty} \Psi_{x_1x_2}(f) \mathrm{e}^{j2\pi f\tau} df, \end{aligned} \quad (4)$$

where $S_{x_1x_2}(f) = E\{X_1(f)X_2^*(f)\}$ is the cross-spectrum, $\Phi(f)$ is a weighting function and

$$\Psi_{x_1x_2}(f) = \Phi(f) S_{x_1x_2}(f) \quad (5)$$

is the generalized cross-spectrum. The GCC TDE may be expressed as:

$$\hat{\tau}_\phi = \arg\max_\tau \psi_{x_1x_2}(\tau). \quad (6)$$

The choice of $\Phi(f)$ is important in practice. The classical cross-correlation (CCC) method is obtained by taking $\Phi(f) = 1$. In the noiseless case, knowing that $X_i(f) = S(f)G_i(f), i = 1, 2$, we have:

$$\Psi_{x_1x_2}(f) = \Psi_{\mathrm{cc}}(f) = G_1(f)E\{|S(f)|^2\}G_2^\star(f). \quad (7)$$

The fact that $\Psi_{\mathrm{cc}}(f)$ depends on the source signal can be problematic for TDE.

The cross-correlation peak can be sharpened by pre-whitening the input signals, i.e. choosing $\Phi(f) = 1/|S_{x_1x_2}(f)|$, which leads to the so-called phase transform (PHAT) method [1, 7]. In the noiseless case, the cross spectrum

$$\Psi_{x_1x_2}(f) = \Psi_{\mathrm{pt}}(f) = G_1(f)G_2^\star(f)/|G_1(f)G_2^\star(f)| \quad (8)$$

depends only on the channel impulse responses and thus can, in general, achieve better performance than CCC.

GCC is simple and easy to implement but will fail when the reverberation becomes important because the simple signal propagation model assumptions are violated.

4. THE PROPOSED METHOD

In this section a completely different approach from GCC is proposed. This new method focuses directly on the channel impulse responses for TDE. First, the principle of this approach is explained and then an algorithm is presented.

4.1. Principle

We assume that the system (room) is linear and time invariant. By following the reverberant model (3) and the fact that $x_1 * g_2 = s * g_1 * g_2 = x_2 * g_1$, in the noiseless case, we have the following relation at time n [8]:

$$\mathbf{x}^T(n)\mathbf{u} = \mathbf{x}_1^T(n)\mathbf{g}_2 - \mathbf{x}_2^T(n)\mathbf{g}_1 = 0 \quad (9)$$

where ^{T} denotes transpose and

$$\begin{aligned}\mathbf{x}_i(n) &= [x_i(n), x_i(n-1), \cdots, x_i(n-M+1)]^T, & (10)\\ \mathbf{g}_i &= [g_{i,0}, g_{i,1}, \cdots, g_{i,M-1}]^T, \ i = 1, 2 & (11)\\ \mathbf{x}(n) &= \left[\mathbf{x}_1^T(n), \mathbf{x}_2^T(n)\right]^T, & (12)\\ \mathbf{u} &= \left[\mathbf{g}_2^T, -\mathbf{g}_1^T\right]^T, & (13)\end{aligned}$$

and M is the length of the impulse responses.

From (9), it can be derived that $\mathbf{R}(n)\mathbf{u} = \mathbf{0}$, where $\mathbf{R}(n) = E\{\mathbf{x}(n)\mathbf{x}^T(n)\}$ is the covariance matrix of the microphone signals $\mathbf{x}(n)$. This implies that the vector $\mathbf{u}$ (containing the two impulse responses) is the eigenvector of the covariance matrix $\mathbf{R}(n)$ corresponding to the eigenvalue equal to 0. Moreover, if the two impulse responses $\mathbf{g}_1$ and $\mathbf{g}_2$ have no common zeros and the autocorrelation matrix of the source signal $s(n)$ is full rank, which is assumed in the rest of this paper, the covariance matrix $\mathbf{R}(n)$ has one and only one eigenvalue equal to 0.

In practice, accurate estimation of the vector $\mathbf{u}$ is not trivial due to the nature of speech, the length of the impulse responses, the background noise, etc. However, for this application we only need to find an efficient way to detect the direct paths of the two impulse responses. In the following, it is explained how this can be done.

4.2. Adaptive Algorithm

In order to efficiently estimate the eigenvector (here $\hat{\mathbf{u}}$) corresponding to the minimum eigenvalue of $\mathbf{R}(n)$, the constrained LMS algorithm [9] is often used. The error signal is

$$e(n) = \frac{\hat{\mathbf{u}}^T(n)\mathbf{x}(n)}{\|\hat{\mathbf{u}}(n)\|}, \quad (14)$$

and the constrained LMS algorithm may be expressed as

$$\hat{\mathbf{u}}(n+1) = \hat{\mathbf{u}}(n) - \mu e(n)\nabla e(n), \quad (15)$$

where μ, the adaptation step, is a positive small constant and

$$\nabla e(n) = \frac{1}{\|\hat{\mathbf{u}}(n)\|}\left[\mathbf{x}(n) - e(n)\frac{\hat{\mathbf{u}}(n)}{\|\hat{\mathbf{u}}(n)\|}\right]. \quad (16)$$

Substituting (14) and (16) in (15) and taking expectation after convergence gives,

$$\mathbf{R}\frac{\hat{\mathbf{u}}(\infty)}{\|\hat{\mathbf{u}}(\infty)\|} = E\{e^2(n)\}\frac{\hat{\mathbf{u}}(\infty)}{\|\hat{\mathbf{u}}(\infty)\|}, \quad (17)$$

which is the desired result: $\hat{\mathbf{u}}$ converges in the mean to the eigenvector of $\mathbf{R}$ corresponding to the smallest eigenvalue $E\{e^2(n)\}$. To avoid roundoff error propagation, normalization is imposed on the vector $\hat{\mathbf{u}}(n+1)$ after each update step. Finally the update equation is given by

$$\hat{\mathbf{u}}(n+1) = \frac{\hat{\mathbf{u}}(n) - \mu e(n)\nabla e(n)}{\|\hat{\mathbf{u}}(n) - \mu e(n)\nabla e(n)\|}. \quad (18)$$

Note that if this normalization is used, then $\|\hat{\mathbf{u}}(n)\|$ (which appears in $e(n)$ and $\nabla e(n)$) can be removed, since we will always have $\|\hat{\mathbf{u}}(n)\| = 1$. If the smallest eigenvalue is equal to zero, which is the case here, the algorithm can be simplified as follows:

$$\begin{aligned}e(n) &= \hat{\mathbf{u}}^T(n)\mathbf{x}(n), & (19)\\ \hat{\mathbf{u}}(n+1) &= \frac{\hat{\mathbf{u}}(n) - \mu e(n)\mathbf{x}(n)}{\|\hat{\mathbf{u}}(n) - \mu e(n)\mathbf{x}(n)\|}. & (20)\end{aligned}$$

Since the goal here is not to accurately estimate the two impulse responses g_1 and g_2 but rather the time delay, only the two direct paths are of interest. In order to take into account negative and positive relative delays, we initialize $\hat{u}_{M/2}(0) = 1$ which will be considered as an estimate of the direct path of g_2 and during

adaption keep it dominant in comparison with the other $M-1$ taps of the first half of $\hat{\mathbf{u}}(n)$ (containing an estimate of the impulse response g_2). A "mirror" effect will appear in the second half of $\hat{\mathbf{u}}(n)$ (containing an estimate of the impuse response $-g_1$): a negative peak will dominate which is an estimate of the direct path of $-g_1$. Thus the relative sample delay will be simply the difference between the indices corresponding to these two peaks.

To take advantage of the FFT, the filter coefficients are updated in the frequency domain in our realtime system using the unconstrained frequency-domain LMS algorithm [10]. The proposed algorithm can be seen as a generalization of the LMS TDE proposed in [11].

5. IMPLEMENTATION

A real-time PC-based TDE system developed in this paper consists of a front-end microphone pair with pre-amps and a 200MHz PentiumePro PC equipped with a standard sound card. Lucent Speech Tracker$^{\mathrm{TM}}$ hypercardioid microphones used in this system can reject noise and reverberation from the sides and rear and allow better signal recording than omni-directional microphones in the half plane in front of the microphones. We used Creative Lab sound blaster AWE32 Board which provides all desired features: 8-bit and 16-bit AD conversion selectable for stereo sampling and playback, variable sample and playback rates from 5kHz to 44.1kHz, etc.

The software is developed with Visual C++$^{\mathrm{TM}}$ under Microsoft Windows operation system. Two application modes are realized:

- realtime: speech is recorded in realtime and active talker direction is immediately available on the PC screen, and
- offline: the system accepts microphone signals from .WAV files, runs the TDE algorithms, and determine the talker location.

Three TDE algorithms have been implemented, namely CCC, PHAT and the proposed algorithm. This system provides a unified and extendable platform for evaluating, simulating, and understanding different TDE algorithms. In addition, a graphical user-friendly interface as shown in Fig.(1) enables users to monitor the input signals continuously, experiment with a variety of TDE parameters, compare consistency of the different TDE algorithms, check their tracking dynamics, and playback the recorded speech.

6. EXPERIMENTAL RESULTS

This section describes the results obtained with developed realtime TDE system and recorded data in the Varechoic chamber. They include consistent and dynamic performance of each implemented TDE algorithm.

6.1. Experimental Setup

The measurements were made in the Varechoic chamber at Bell Labs [12]. A diagram of the floor plan layout is shown in Fig.2 with the position of the sources (loudspeaker at different positions) and two microphones which are apart about 37.25in (95cm). Three different panel configurations were selected and the corresponding 60dB reverberation time in the $400-1600$Hz band is 150ms, 250ms and 740ms, respectively.

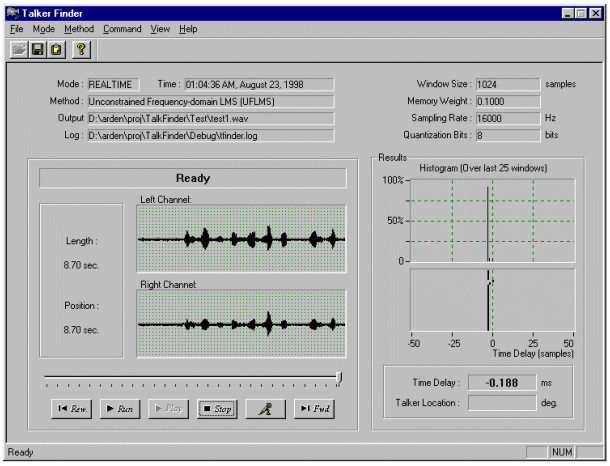

Figure 1: The system graphical user interface.

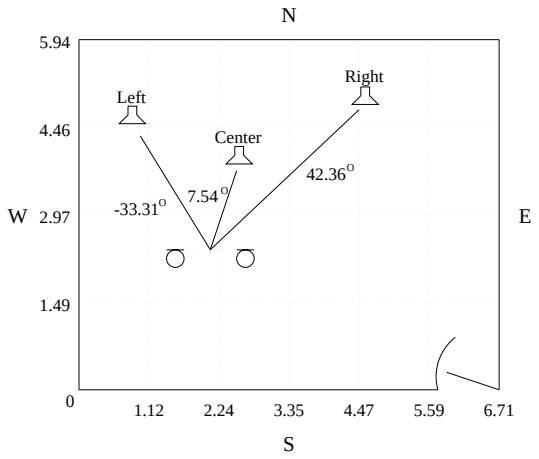

Figure 2: Varechoic chamber floor plan (coordinate values measured in meters) with the position of the two microphones and the sources.

The microphone signals are sampled simultaneously at a rate of 16kHz which causes a maximum $1/16$ms (about 2cm) synchronization error for TDE. Each recording was about 5s long. In our measurement, PHAT and CCC use a 64ms Kaiser window for the analysis frame. For the proposed algorithms, the step size is chosen to be $\mu = 0.003$ and the length of the adaption vector $\hat{\mathbf{u}}(n)$ is taken as $L = 2M = 512$. The power of the two microphone signals $X_i(f,n)$, $i = 1, 2$, was estimated in the frequency-domain as follows:

$$P_i(f,n) = \gamma P_i(f,n-1) + (1-\gamma)|X_i(f,n)|^2, \qquad (21)$$

with $\gamma = 0.1$. We initialized the algorithm with $P_i(f,0) = 2L\sigma_{x_i}^2$ ($\sigma_{x_i}^2$ is the average power of x_i).

6.2. Consistency

Figure 3 shows histograms of TDE with the source located on the right in Figure 2. Many more consistency testings have been made and are detailed in [13], but only one set is given here due to space

restrictions. It can be seen that the new method is more consistent and more accurate than PHAT and CCC. Notably with a 740ms reverberation time, all methods fail except for the proposed algorithm.

6.3. Tracking Dynamics

The realtime system has been tested in an office room at Bell Labs in the presence of background noise and reverberation at a moderate level to examine the tracking dynamics of different TDE algorithms. For PHAT and CCC, the tracking delay is negligible but these algorithms are very sensitive to the noise. Some nonspeech events, such as door opening and key swinging, would change the time delay estimate, which is not desirable for camera pointing or beamformer steering. However, the proposed algorithm trades off the tracking delay with the system sensitivity and consistency. It converges to a good time delay estimate in less than 250ms which is tolerable in most TDE applications. The system takes at most 1 seconds to correctly locate an active speaker.

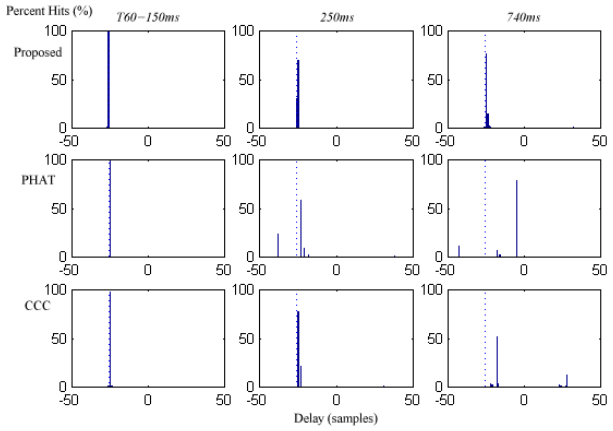

Figure 3: Histograms of TDE with a speech signal located on the right in Fig. 2. The first, second, and third columns correspond respectively to a reverberation time of 150ms, 250ms, and 740ms. The first, second, and third rows correspond respectively to the TDE by the proposed algorithm, PHAT, and CCC. The true delay is plotted with a dotted line.

7. CONCLUSION

In this paper, a new and simple approach to time delay estimation has been proposed and a low-cost realtime PC-based acoustic source localization system was presented. The method consists of detecting the direct paths of the two impulse responses between the source signal and the microphones that are estimated in the eigenvector corresponding to the smallest eigenvalue of the covariance matrix of the microphone signals. The system is capable of determining the location of an active talker in realtime and provides a unified and extendable platform with graphical user interface for TDE algorithm testing. The proposed algorithm offers several advantages. It is easy to implement for realtime operation and, in comparison with other methods, seems to be more effective in a reverberant environment and much more accurate.

8. ACKNOWLEDGMENTS

The authors would like to thank Jens Meyer for recording the data in the Varechoic chamber, Dennis R. Morgan for his valuable suggestions on the user interface design and Tomas Gaensler for conducting user tests and giving helpful comments for improvement. We also thank Kathleen L. Shipley for her guidance in Windows programming, Weicong Wang for his discussions while code debuging, and Robert A. Kubli for lending audio recording facilities and providing technical support in their setup.

9. REFERENCES

[1] 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, no. 4, pp. 320-327, Aug. 1976.

[2] M. S. Brandstein, "A pitch-based approach to time-delay estimation of reverberant speech," in *Proc. IEEE ASSP Workshop Appls. Signal Processing Audio Acoustics*, 1997.

[3] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. IEEE ASSP Workshop Appls. Signal Processing Audio Acoustics*, 1997.

[4] B. Champagne, S. Bédard, and A. Stéphenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Trans. Speech Audio Processing*, vol. 4, no. 2, pp. 148-152, Mar. 1996.

[5] A. Stéphenne and B. Champagne, "Cepstral prefiltering for time delay estimation in reverberant environments," in *Proc. IEEE ICASSP*, 1995, pp. 3055-3058.

[6] M. Bodden, "Modeling human sound-source localization and the cocktail-party-effect", Acta Acoustica 1, pp. 43-55, 1993.

[7] M. Omologo and P. Svaizer, "Acoustic source location in noisy and reverberant environment using CSP analysis," in *Proc. IEEE ICASSP*, 1996, pp. 921-924.

[8] J. Benesty, F. Amand, A. Gilloire, and Y. Grenier, "Adaptive filtering algorithms for stereophonic acoustic echo cancellation," in *Proc. IEEE ICASSP*, 1995, pp. 3099-3102.

[9] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. of the IEEE*, vol. 60, no. 8, pp. 926-935, Aug. 1972.

[10] D. Mansour and A. H. Gray, JR., "Unconstrained frequency-domain adaptive filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, no. 5, pp. 726-734, Oct. 1982.

[11] D. H. Youn, N. Ahmed, and G. C. Carter, "On using the LMS algorithm for time delay estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, no. 5, pp. 798-801, Oct. 1982.

[12] W. C. Ward, G. W. Elko, R. A. Kubli, and W. C. McDougald, "The new Varechoic chamber at AT&T Bell Labs," in *Proc. Wallance Clement Sabine Centennial Symposium*, 1994, pp. 343-346.

[13] J. Benesty, " Adaptive Eigenvalue Decomposition Algorithm for Passive Acoustic Source Localization", *Bell Labs Tech. Memo.* , 1998.