MICROPHONE ARRAY SOURCE LOCALIZATION USING REALIZABLE DELAY VECTORS

Scott M. Griebel and Michael S. Brandstein

Division of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138
griebel@fas.harvard.edu, msb@hrl.harvard.edu

ABSTRACT

This paper presents an alternative approach to acoustic source localization which modifies the traditional two-step localization procedure to not require explicit time-delay estimates. Instead, the cross-correlation functions derived from various microphone pairs are simultaneously maximized over a set of potential delay combinations consistent with candidate source locations. The result is a procedure that combines the advantages offered by the PHAT weighting (or any reasonable cross-correlation type function) and a more robust localization procedure without dramatically increasing computational load. Simulations are performed across a range of reverberation conditions to illustrate the utility of the proposed method relative to conventional GCC filtering approaches and a more modern eigenvalue-based technique.

1. INTRODUCTION

An essential requirement of microphone array systems is the ability to locate and track one or more speech sources. For speech enhancement applications, accurate knowledge of the desired talker and interference locations is necessary to effectively steer the array, enhancing a given source while simultaneously attenuating those deemed undesirable. Location data may be used as a guide for discriminating individual speakers in a multi-source scenario. With this information available, it would then be possible to automatically focus upon and follow a given source on an extended basis. Of particular interest lately, is the application of the speaker location estimates for aiming a camera or series of cameras in a video-conferencing system.

While beamforming has been used extensively in speech-array applications for voice capture it has rarely been applied to the talker localization problem. The task of computing the steered output power for an appropriately dense set of candidate locations is computationally burdensome and highly dependent upon the spectral content of the source signal. Primarily because of computational practicality and reasonable performance under amicable conditions, the bulk of passive talker localization systems in use today utilize a two-step procedure. Time delay estimates (TDE's) of the speech signal relative to pairs of spatially separated microphones are obtained and then applied to generate a source location estimate. A number of variations on this principle have been developed. They differ considerably in their method of derivation, the extent of their applicability (2-D vs. 3-D, near source vs. distant source, etc.), and their means of solution. See [1] for details.

Accurate and robust time-delay estimation is the key to the effectiveness of localizers within this genre. The two major sources of signal degradation which complicate this estimation problem are environmental background noise and channel multipath due to room reverberations. The noise-alone case has been addressed at length and estimating the time delay from the peak of a SNR-weighted version of the Generalized Cross-Correlation (GCC) function [2] has been shown to be optimal under certain noise conditions. The problem becomes more challenging once room reverberations rise above minimal levels and the SNR-weighted GCC quickly becomes unreliable [3, 4]. A basic approach to dealing with multipath channel distortions in this context has been to make the GCC function more robust by deemphasizing the frequency-dependent weightings. The Phase Transform (PHAT) [2] is one extreme of this procedure which has received considerable attention as the basis of speech source localization systems [5, 6, 7]. By weighting each component of the cross-spectrum phase equally, the PHAT deemphasizes signal content in favor of the desired room response in the resulting GCC function. The PHAT GCC-based time-delay estimator has been shown to be fairly precise in low to moderate reverberation conditions. However, with increasing reverberation levels it is subject to anomalous behavior in which a multipath reflection is confused with the direct path delay. Although the resulting cross-correlation functions often do have local maxima at the true time delay, they are not global maxima, and therefore lead to erroneous TDE's. This leads to one or more of the time delay estimates for a microphone array to be inaccurate and detrimentally affects the second step in the localization procedure.

Motivated by the preceding observation this paper explores an alternative approach to acoustic source localization which modifies the two-step localization procedure to not require explicit time-delay estimates. Instead, the entire GCC functions derived from various microphone pairs are simultaneously maximized over a set of potential delay combinations consistent with candidate source locations. The result is a procedure that combines the advantages offered by the PHAT weighting (or any reasonable cross-correlation type function) and a more robust localization procedure without dramatically increasing computational load. These results may be viewed as a special (and computationally practical) case of the SRP-PHAT algorithm described in [8]. Another less general technique was presented in [9] which assumes that all microphone pairs are centered around the same location. This theoretically leads to a peak in the sum of the individual cross-power spectra at the correct time delay.

This work was funded by National Science Foundation CAREER grant CCR-9983839.

The following section overviews the traditional method for TDE-based localization. Section 3 details the proposed localization method while Section 4 presents the results of simulations validating its utility. The paper concludes with some discussion.

2. TRADITIONAL TDE-BASED LOCALIZATION

The signals received at two microphones, $x_1(t)$ and $x_2(t)$, may be modeled as:

$$\begin{aligned} x_1(t) &= h_1(t) * s(t) + n_1(t) \qquad 0 \le t \le T \\ x_2(t) &= h_2(t) * s(t - \tau_{12}) + n_2(t) \end{aligned}$$

where τ_{12} is the relative signal delay of interest, $h_1(t)$ and $h_2(t)$ represent the impulse responses of the reverberant channels, $s(t)$ is the speech signal, $n_1(t)$ and $n_2(t)$ correspond to additive noise uncorrelated with $s(t)$, and $*$ denotes linear convolution.

The Generalized Cross-Correlation (GCC) function for a given time-lag, $R_{12}(\tau)$, is calculated as the inverse Fourier transform of the received signal cross-spectrum, $X_1(\omega)X_2'(\omega)$, scaled by a weighting function, $W(\omega)$:

$$R_{12}(\tau) = \int_{-\infty}^{\infty} W(\omega) X_1(\omega) X_2'(\omega) e^{j\omega\tau} d\omega. \qquad (1)$$

The delay estimate, $\hat{\tau}_{12}$, corresponds to the lag value which maximizes $R_{12}(\tau)$. The standard cross-correlation function is obtained when the weighting function is uniform, i.e. $W_{XC}(\omega) = 1$. The resulting correlation function will be called $R^{XC}(\tau)$. The Phase Transform employs the weighting function:

$$W_{PHAT}(\omega) = |X_1(\omega)X_2'(\omega)|^{-1} \qquad (2)$$

to eliminate the influence of the spectral magnitudes and produce a GCC function dependent entirely on the phase of the cross-spectrum. The resulting correlation function will be called $R^{PHAT}(\tau)$. Typically, a parabolic fit is performed about the peak in $R^{XC}(\tau)$ or $R^{PHAT}(\tau)$ to achieve sub-sample resolution.

Given a set of M TDE's derived from the signals received at multiple pairs of microphones, the problem remains as how to best estimate the true source location, $\mathbf{q_s}$. Ideally, the estimate will be an element of the intersection of all the potential source loci. In practice, however, for more than two pairs of sensors this intersection is, in general, the empty set. With no ideal solution available, the source location must be estimated as the point in space which best *fits* the sensor-TDE data or more specifically, minimizes an error criterion that is a function of the given data and a hypothesized source location. If the TDE's at each microphone pair are assumed to be independently corrupted by zero-mean additive white Gaussian noise of equal variance then the ML location estimate can be shown to be the position which minimizes the least squares error criterion [1]:

$$E_{LS}(\mathbf{q}) = \sum_{i=1}^{M} (\hat{\tau}_i - T(\{\mathbf{p_{i1}}, \mathbf{p_{i2}}\}, \mathbf{q}))^2 \qquad (3)$$

where $\hat{\tau}_i$ is the TDE associated with the i^{th} pair of microphones with spatial coordinates denoted by the 3-element vectors, $\mathbf{p_{i1}}$ and $\mathbf{p_{i2}}$, and $T(\{\mathbf{p_{i1}}, \mathbf{p_{i2}}\}, \mathbf{q})$ is the true TDE for a signal at source location $\mathbf{q}$ relative to the i^{th} microphone pair.

The source location derived in this way will be referred as $\hat{\mathbf{q}}_{LS}$. Note that this procedure involves the optimization of a nonlinear function and necessitates the use of search methods. From a robustness and complexity standpoint this search is significantly simpler than that required by beamsteering approaches. Despite this, a number of closed form approximations to this method exist in the literature; [10, 11, 12] are typical.

3. LOCALIZATION USING REALIZABLE DELAY VECTORS

Figure 1 illustrates the shortcomings of the traditional TDE-based localization scheme. It shows a set of bearing lines derived from the TDE's obtained when using the PHAT GCC independently for an array of 8 microphones which are indicated by an $\times$. The TDE's were estimated for each pair of neighboring microphones. For this example, the reverberation time was 400ms and there was additive white noise with a SNR of 15dB. The source, indicated by a square, was located at $(x, y) = (1.85, 2.75)$. As is apparent in the plot, two bearing lines appear to be focusing on first order wall reflections rather than the desired source. Without judicious pruning of these two delay estimates in the least squares error criterion, the resulting traditional source location estimate, $\hat{\mathbf{q}}_{LS}$, will be significantly biased away from its true location.

Figure 2 shows the PHAT GCC function for the rightmost pair of microphones in Figure 1. The true delay is -3.01 samples, but the peak in the PHAT GCC occurs at -11.2 samples. This leads to a bearing line estimate that is well to the left of the true bearing line. A local maximum does occur in the PHAT GCC at the true time delay, but it is not the global maximum and is therefore not chosen as the TDE.

As the above plot suggests, reducing the information in the cross-correlation function down to a single parameter, the TDE, prior to estimating the source location is problematic. Rather than selectively culling or weighting TDE's in a error criterion, we have found that it is advantageous to work directly with the GCC functions themselves to generate a delay vector, $\boldsymbol{\tau_{con}}$, which is realizable in the sense that it is consistent with a physical source location.

The time delay vector is estimated from:

$$\hat{\tau}_{con} = \max_{\tau \in T_{con}} \sum_{i=1}^{M} \phi_i(\tau_i)^{\alpha} \qquad (4)$$

where M is the number of microphone pairs, ϕ_i is the cross-correlation function for the i^{th} pair, τ is an M-element vector belonging to the set of delay vectors consistent with potential source locations, T_{con}, and α is an exponential weighting term (typically $\alpha = 1, 2$). Computed in this way, the corresponding source location estimate is simply the spatial location consistent with $\hat{\tau}_{con}$.

With this structure, any cross-correlation like function may be substituted for $\phi(\tau)$. For example, the results to follow will consider the use of standard cross-correlation and PHAT GCC methods, $\phi(\tau) = R^{XC}(\tau)$ and $\phi(\tau) = R^{PHAT}(\tau)$, respectively, as well as an appropriate function derived from eigenvalue decomposition-based time delay estimation algorithm presented by Benesty in [13]. Also note that since discrete-time functions are involved, it is necessary to interpolate $\phi(\tau)$ to obtain intra-sample resolution.

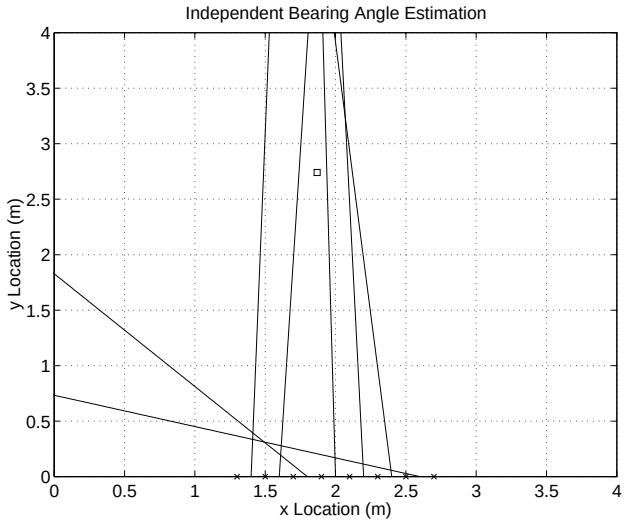


Figure 1: Bearing angle estimates derived from independent PHAT GCC-based time delay estimates. The true source location is indicated by □ and the 8 microphones are indicated by ×.

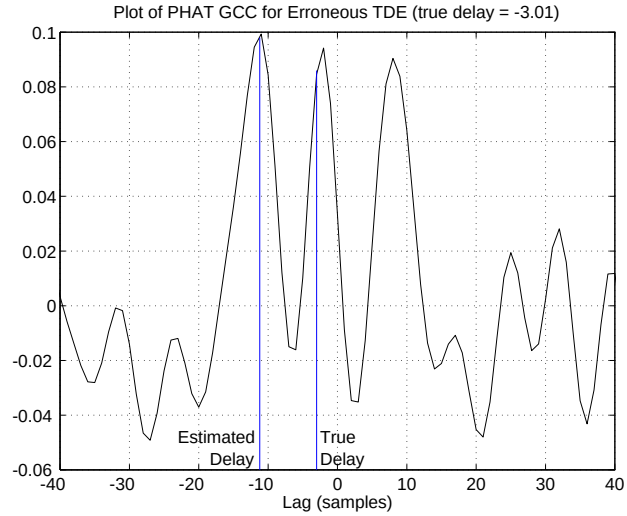


Figure 2: Plot of the PHAT GCC for the rightmost microphone pair in Figure 1. The true delay is -3.01 samples, while the estimated delay is -11.2 samples.

4. RESULTS

To evaluate the utility of the proposed localization scheme, simulations were performed in a 4m × 4m × 3m room using using 10 randomly generated source locations and reverberation times ranging from 0 to 1000ms. 8 microphones arranged as two 4-element linear arrays (i.e. a 4×2 grid) with .2m spacing in both dimensions were positioned on one wall of the enclosure. In all the simulations to follow, cross-correlation functions were computed only for neighboring microphones (i.e., those separated by .2m), yielding a value of $M = 10$. Impulse responses were generated using the image model of [14]. White noise resulting in an SNR of 15dB was added to each microphone signal to make the test more challenging and realistic.

Figure 3 offers some qualitative comparisons. The least-squares error (LS) and consistency criterion (CON), equations (3) and (4) respectively, for a source in the farfield of the array and a reverberation level of 400ms are plotted for three different functions: standard cross-correlations (XC), Benesty's algorithm (BEN), and the PHAT GCC (PHAT). The consistency plots in the righthand column all show a superior ability to discriminate the source localization relative to their LS counterparts in the lefthand column. The PHAT-CON case exhibits the best performance. Figure 4 repeats this experiment with 2 simultaneous talkers. As expected the LS criterion is unable to deal with the multi-talker scenario and the extrema in the corresponding plots tends to be somewhere in the middle of the two desired locations. The XC-CON and BEN-CON represent only mild improvements while the PHAT consistent estimation method clearly exhibits two distinct direction of arrival locations.

Figure 5 presents an analysis of the 6 localization combinations evaluated for the 10 random source locations and the range of reverberation levels. In order to compare these techniques, a location estimate was considered correct if the resulting bearing angle relative to the array center was less than 5°. Note that the consistent versions of each algorithm always performed at least as well as their LS counterparts. Also note that the XC function does not gain much from enforcing consistency. This suggests that when errant time delays are observed, the correlation peaks corresponding to the true delay were very small. However, the PHAT and Benesty techniques show a dramatic improvement when enforcing consistency, with the Benesty algorithm peforming better for lower reverberation times (0-200ms) and the PHAT algorithm performing better for higher reverberation times (300-1000ms).

5. CONCLUSIONS

As these simulations suggest, the traditional two-step localization procedure clearly suffers from its intermediate use of explicit TDE's. The consistency-based localization scheme more effectively exploits the information available in the cross-correlation function by incorporating knowledge of the physical constraints dictated by the source-microphone geometry.

The proposed method may be employed with any cross-correlation like function. Here, the PHAT GCC and Benesty functions were focused upon because of their proven utility for the TDE problem. The advantages of the PHAT weighting, namely its deemphasis of signal content and robustness to reverberation, are also apparent (and enhanced) in this application. The function derived from the Benesty algorithm, while not as effective at moderate to high reverberation levels, does offer advantage under the low reverberation conditions where the additive noise is the dominant source of degradation.

From a computational standpoint, the proposed localization scheme represents a middle ground between the hyper-efficient closed-form localizers and the burdensome beamsteering methods. As the plots in Figures 3 and 4, the search space is very amenable to iterative approaches and the consistency criterion in (4) is straightforward to calculate.

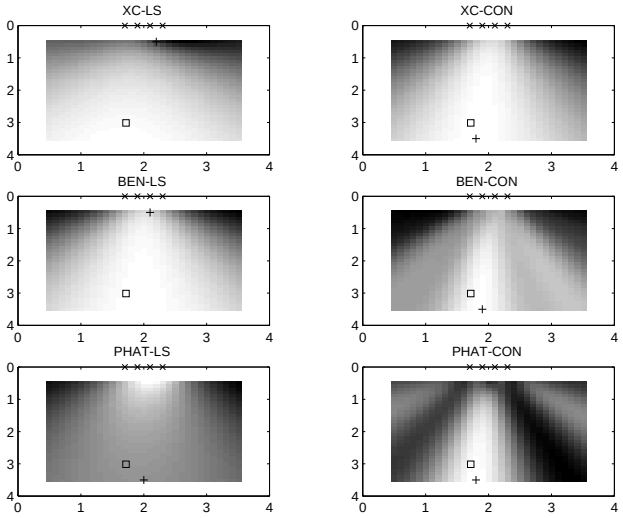


Figure 3: Qualitative plots of the LS and consistency criteria typical of a farfield source for the three cross-correlation like functions considered. The source location is indicated by □ and the 8 microphones (two rows of 4) by ×.

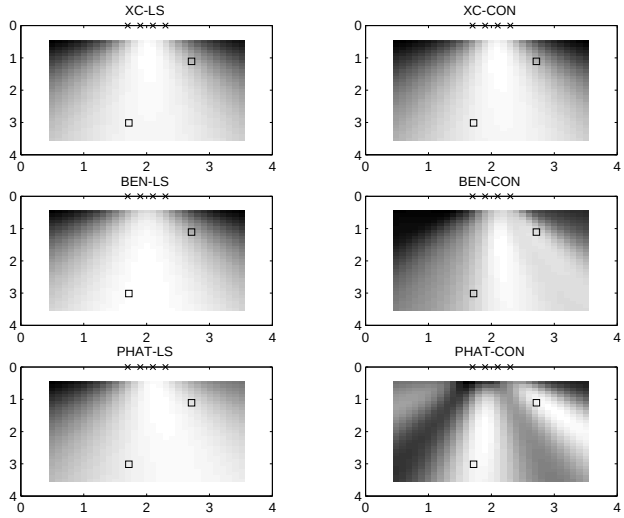


Figure 4: The qualitative plots of Figure 3 repeated for 2 simultaneous sources.

6. REFERENCES

[1] M. Brandstein and H. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer, Speech, and Language*, vol. 11, no. 2, pp. 91–126, April 1997.

[2] 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, no. 4, pp. 320–327, August 1976.

[3] S. Bédard, B. Champagne, and A. Stéphenne, "Effects of room reverberation on time-delay estimation performance," in *Proceedings of ICASSP94*. IEEE, 1994, pp. II–261– II–264.

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

[5] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event localization," *IEEE Trans. Speech Audio Proc.*, vol. 5, no. 3, pp. 288–292, May 1997.

[6] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proceedings of ICASSP97*, Munich, Germany, April 20-24 1997, IEEE, pp. 187–190.

[7] P. Svaizer, M. Matassoni, and M. Omologo, "Acoustic source location in a three-dimensional space using crosspower spectrum phase," in *Proceedings of ICASSP97*, Munich, Germany, April 20-24 1997, IEEE, pp. 231–234.

[8] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Springer Verlag, 2001.

[9] T. Nishiura, T. Yamada, S. Nakamura, and K. Shikano, "Localization of multiple sound sources based on a csp analysis with a microphone array," in *Proceedings of ICASSP2000*. IEEE, 2000, pp. II–1053– II–1056.

[10] M. Brandstein, J. Adcock, and H. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Trans. Speech Audio Proc.*, vol. 5, no. 1, pp. 45–50, January 1997.

[11] Y. Huang, J. Benesty, and G. Elko, "Microphone arrays for video camera steering," in *Acoustic Signal Processing for Telecommunication*, S. Gay and J. Benesty, Eds., pp. 239–259. Kluwer, 2000.

[12] J. Abel and J. Smith, "Source range and depth estimation from multipath range difference measurements," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, no. 8, pp. 1157–1165, August 1989.

[13] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Am.*, vol. 107, no. 1, pp. 384–391, 2000.

[14] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, April 1979.

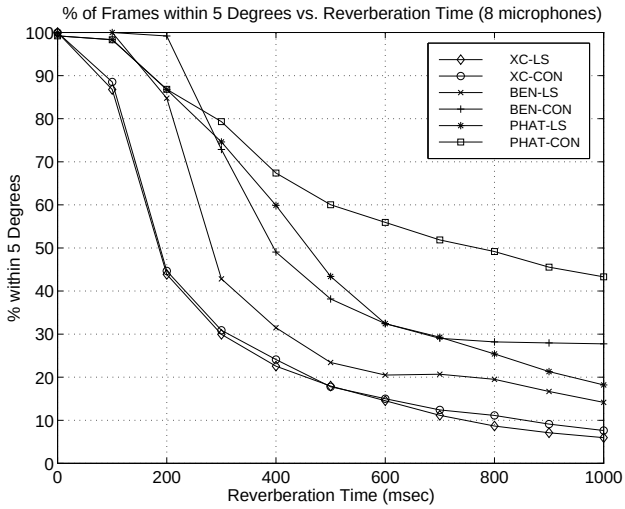


Figure 5: Analysis of the LS (XC-LS, BEN-LS, PHAT-LS) and consistency-based(XC-CON, BEN-CON, PHAT-CON) locators for various reverberation levels and 15dB SNR additive white noise