MAXIMUM LIKELIHOOD SOUND SOURCE LOCALIZATION FOR MULTIPLE DIRECTIONAL MICROPHONES

Cha Zhang, Zhengyou Zhang and Dinei Florêncio

Microsoft Research, One Microsoft Way, Redmond, 98052
Email: {chazhang,zhang,dinei}@microsoft.com

ABSTRACT

This paper presents a maximum likelihood (ML) framework for multi-microphone sound source localization (SSL). Besides deriving the framework, we focus on making the connection and contrast between the ML-based algorithm and popular steered response power (SRP) SSL algorithms such as phase transform (SRP-PHAT). We also show under our ML framework how challenging conditions such as directional microphone arrays and reverberations can be handled. The computational cost of our method is low – similar to SRP-PHAT. The effectiveness of the proposed method is shown on a large dataset with 99 real-world audio sequences recorded by directional circular microphone arrays in over 50 different meeting rooms.

Index Terms— Sound source localization, maximum likelihood, directional circular microphone arrays

1. INTRODUCTION

Sound source localization (SSL) using microphone arrays has been an active research topic since the early 1990's [1]. It has found many important applications such as human-computer interaction [2, 3] and intelligent rooms [4, 5]. A large number of SSL algorithms have been proposed in literature, with varying degrees of accuracy and computational complexity.

For broadband acoustic source localization applications such as teleconferencing, a number of SSL techniques are popular, including steered-beamformer (SB) based, high-resolution spectral estimation based, time delay of arrival (TDOA) based [1], and learning based [6]. Among them, the TDOA based approaches have received extensive investigation [1, 4, 7, 8, 9, 10]. In this paper, we derive a TDOA based maximum likelihood (ML) framework for multi-microphone sound source localization. While this is not the first time ML estimation is applied for SSL [11, 12, 13, 14], our derivation allows us to build a connection between the ML based SSL and the popular SRP based SSL algorithms, which are known to work extremely well in practical environments [15, 4, 16] and have very low computational cost. We demonstrate within the ML framework how reverberation can be treated by introducing an additional term during noise modeling, and how the different gains of microphones (e.g., when directional microphones are used in the array) can be compensated from the received signal and the noise model. The effectiveness of the proposed method is shown on a real-world dataset containing 99 audio sequences recorded by directional circular microphone arrays in over 50 meeting rooms.

The rest of the paper is organized as follows. We review a number of popular SSL approaches in Section 2. The ML framework is derived in Section 3. Relationship between the ML SSL algorithm and various existing approaches is discussed in Section 4. Experiments and conclusions are given in Section 5 and 6, respectively.

2. REVIEW OF EXISTING APPROACHES

Consider an array of P microphones. Given a source signal $s(t)$, the signals received at these microphones can be modeled as [9, 16]:

$$x_i(t) = \alpha_i s(t - \tau_i) + h_i(t) \otimes s(t) + n_i(t), \quad (1)$$

where $i = 1, \cdots, P$ is the index of the microphones, τ_i is the time of propagation from the source location to the i^{th} microphone; α_i is a gain factor that includes the propagation energy decay of the signal, the gain of the corresponding microphone, the directionality of the source and the microphone, etc; $n_i(t)$ is the noise sensed by the i^{th} microphone; $h_i(t) \otimes s(t)$ represents the convolution between the environmental response function and the source signal, often referred as the *reverberation*. In many existing SSL approaches [17, 1, 8, 10], the reverberation term was ignored for simplicity. In the frequency domain, we can rewrite the above model as:

$$X_i(\omega) = \alpha_i(\omega) S(\omega) e^{-j\omega\tau_i} + H_i(\omega) S(\omega) + N_i(\omega), \quad (2)$$

where we also allow the α_i to vary with frequency.

The most straightforward SSL algorithm, is to take each pair of the microphones and compute their cross-correlation function. For instance, the correlation between the signals received at microphone i and k can be computed in the frequency domain as:

$$R_{ik}(\tau) = \int X_i(\omega) X_k^*(\omega) e^{j\omega\tau} d\omega, \quad (3)$$

where * represents complex conjugate. The τ that maximizes the above correlation is the estimated time delay between the two signals. When more than two microphones are considered, one can sum over all possible pairs of microphones and have:

$$\mathcal{R}(\mathbf{s}) = \sum_{i=1}^{P} \sum_{k=1}^{P} \int X_i(\omega) X_k^*(\omega) e^{j\omega(\tau_i - \tau_k)} d\omega, \quad (4)$$

$$= \int \Big| \sum_{i=1}^{P} X_i(\omega) e^{j\omega\tau_i} \Big|^2 d\omega, \quad (5)$$

The common practice is to maximize the above correlation through hypothesis testing, where $\mathbf{s}$ is the hypothesized source location, which determines the τ_i's on the right. Eq. (5) is also known as the steered response power (SRP) of the microphone array.

To address the reverberation and noise that may affect the SSL accuracy, researchers found that adding a weighting function in front of the correlation can greatly help. Eq. (4) is thus rewritten as:

$$\mathcal{R}(\mathbf{s}) = \sum_{i=1}^{P} \sum_{k=1}^{P} \int \Psi_{ik}(\omega) X_i(\omega) X_k^*(\omega) e^{j\omega(\tau_i - \tau_k)} d\omega, \quad (6)$$

A number of weighting functions have been investigated in literature [17]. Among them, the heuristic-based PHAT weighting [17] defined as:

$$\Psi_{ik}(\omega) = \frac{1}{|X_i(\omega)X_k^*(\omega)|} = \frac{1}{|X_i(\omega)||X_k(\omega)|} \tag{7}$$

has been found to perform very well under realistic acoustical conditions [4, 16]. Inserting Eq. (7) into Eq. (6), one gets:

$$\mathcal{R}(\mathbf{s}) = \int \Big| \sum_{i=1}^{P} \frac{X_i(\omega)e^{j\omega\tau_i}}{|X_i(\omega)|} \Big|^2 d\omega, \tag{8}$$

This algorithm is called SRP-PHAT [15]. Note SRP-PHAT is very efficient to compute, because the number of weighting and summations drops from P^2 in Eq. (6) to P.

A more theoretically-sound weighting function is the maximum likelihood (ML) formulation given by Brandstein et al [1], assuming high signal to noise ratio and no reverberation. The weighting function of a microphone pair is defined as:

$$\Psi_{ij}(\omega) = \frac{|X_i(\omega)||X_j(\omega)|}{|N_i(\omega)|^2|X_j(\omega)|^2 + |N_j(\omega)|^2|X_i(\omega)|^2}. \tag{9}$$

Eq. (9) can be inserted into Eq. (6) to obtain a ML based algorithm. This algorithm is known to be robust to noises, but its performance in real-world applications is relatively poor, because reverberation is not modeled during its derivation. In [16], Rui and Florêncio improved the algorithm by treating the reverberation as another type of noise, same as [4]:

$$|N_i^c(\omega)|^2 = \gamma|X_i(\omega)|^2 + (1-\gamma)|N_i(\omega)|^2, \tag{10}$$

where $N_i^c(\omega)$ is the combined noise or total noise. Eq. (10) is then plugged into Eq. (9) (replacing $N_i(\omega)$ with $N_i^c(\omega)$) to obtain the new weighting function. Their follow-up work [18] used some further approximation and gave:

$$\mathcal{R}(\mathbf{s}) = \int \Big| \sum_{i=1}^{P} \frac{X_i(\omega)e^{j\omega\tau_i}}{\gamma|X_i(\omega)| + (1-\gamma)|N_i(\omega)|} \Big|^2 d\omega, \tag{11}$$

whose computational efficiency is close to SRP-PHAT.

Note, however, that algorithms derived from Eq. (9) are not true ML algorithms for multiple microphones. This is because the optimal weight in Eq. (9) was derived only for two microphones. When more than 2 microphones are used, the adoption of Eq. (6) assumes that pairs of microphones are independent and their likelihood can be multiplied together, which is questionable. In the next section, a true ML algorithm will be developed for the case of multiple microphones. And we will show the connection between the ML algorithm and the existing algorithms in Section 4.

3. THE PROPOSED FRAMEWORK

Let us start by rewriting Eq. (2) into a vector form:

$$\mathbf{X}(\omega) = S(\omega)\mathbf{G}(\omega) + S(\omega)\mathbf{H}(\omega) + \mathbf{N}(\omega), \tag{12}$$

where

$$\begin{aligned}
\mathbf{X}(\omega) &= [X_1(\omega), \cdots, X_P(\omega)]^T, \\
\mathbf{G}(\omega) &= [\alpha_1(\omega)e^{-j\omega\tau_1}, \cdots, \alpha_P(\omega)e^{-j\omega\tau_P}]^T, \\
\mathbf{H}(\omega) &= [H_1(\omega), \cdots, H_P(\omega)]^T, \\
\mathbf{N}(\omega) &= [N_1(\omega), \cdots, N_P(\omega)]^T.
\end{aligned}$$

Among the variables, $\mathbf{X}(\omega)$ represents the received signals, hence it is known. $\mathbf{G}(\omega)$ can be estimated or hypothesized during the SSL process, which will be detailed later. The reverberation term $S(\omega)\mathbf{H}(\omega)$ is unknown, and we will treat it as another type of noise.

To make the above model mathematically tractable, we assume the combined total noise,

$$\mathbf{N}^c(\omega) = S(\omega)\mathbf{H}(\omega) + \mathbf{N}(\omega), \tag{13}$$

follows a zero-mean, independent between frequencies, joint Gaussian distribution, i.e.,

$$p(\mathbf{N}^c(\omega)) = \rho \exp\Big\{ -\frac{1}{2}[\mathbf{N}^c(\omega)]^H \mathbf{Q}^{-1}(\omega)\mathbf{N}^c(\omega) \Big\}, \tag{14}$$

where ρ is some constant; superscript H represents Hermitian transpose, $\mathbf{Q}(\omega)$ is the covariance matrix, which can be estimated by:

$$\begin{aligned}
\mathbf{Q}(\omega) &= E\{\mathbf{N}^c(\omega)[\mathbf{N}^c(\omega)]^H\} \\
&= E\{\mathbf{N}(\omega)\mathbf{N}^H(\omega)\} + |S(\omega)|^2 E\{\mathbf{H}(\omega)\mathbf{H}^H(\omega)\}
\end{aligned} \tag{15}$$

Here we assume the noise and the reverberation are uncorrelated. The first term in Eq. (15) can be directly estimated from the silence periods of the acoustical signals:

$$E(N_i(\omega)N_j^*(\omega)) = \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} N_{ik}(\omega)N_{jk}^*(\omega), \tag{16}$$

where k is the index of audio frames that are silent. Note the background noises received at different microphones may be correlated, such as the ones generated by computer fans in the room. If we believe the noises are independent at different microphones, we can simplify the first term of Eq. (15) further as a diagonal matrix:

$$E\{\mathbf{N}(\omega)\mathbf{N}^H(\omega)\} = \text{diag}(E\{|N_1(\omega)|^2\}, \cdots, E\{|N_P(\omega)|^2\}). \tag{17}$$

The second term in Eq. (15) is related to reverberation. It is generally unknown. As an approximation, we assume it is diagonal:

$$|S(\omega)|^2 E\{\mathbf{H}(\omega)\mathbf{H}^H(\omega)\} \approx \text{diag}(\lambda_1, \cdots, \lambda_P), \tag{18}$$

with the i^{th} diagonal element as:

$$\begin{aligned}
\lambda_i &= E\{|H_i(\omega)|^2|S(\omega)|^2\} \\
&\approx \gamma(|X_i(\omega)|^2 - E\{|N_i(\omega)|^2\})
\end{aligned} \tag{19}$$

where $0 < \gamma < 1$ is an empirical parameter. Eq. (19) assumes that the reverberation energy is a portion of the difference between the total received signal energy and the environmental noise energy. The same assumption was used in Eq. (10) [4, 16]. Note again that Eq. (18) is an approximation, because normally the reverberation signals received at different microphones are correlated, and the matrix should have non-zero off-diagonal elements. Unfortunately, it is generally very difficult to estimate the actual reverberation signals or these off-diagonal elements in practice. In the following analysis, we will use $\mathbf{Q}(\omega)$ to represent the noise covariance matrix, hence the derivation is applicable even when it does contain non-zero off-diagonal elements.

When the covariance matrix $\mathbf{Q}(\omega)$ can be estimated from known signals, the likelihood of the received signals can be written as:

$$p(\mathbf{X}|S, \mathbf{G}, \mathbf{Q}) = \prod_{\omega} p(\mathbf{X}(\omega)|S(\omega), \mathbf{G}(\omega), \mathbf{Q}(\omega)), \tag{20}$$

where

$$p(\mathbf{X}(\omega)|S(\omega),\mathbf{G}(\omega),\mathbf{Q}(\omega)) = \rho\exp\big\{-J(\omega)/2\big\}, \quad (21)$$

$$J(\omega) = [\mathbf{X}(\omega) - S(\omega)\mathbf{G}(\omega)]^H\mathbf{Q}^{-1}(\omega)[\mathbf{X}(\omega) - S(\omega)\mathbf{G}(\omega)]. \quad (22)$$

The goal of the proposed sound source localization is thus to maximize the above likelihood, given the observations $\mathbf{X}(\omega)$, gain matrix $\mathbf{G}(\omega)$ and noise covariance matrix $\mathbf{Q}(\omega)$. Note the gain matrix $\mathbf{G}(\omega)$ requires information about where the sound source comes from, hence the optimization is usually solved through hypothesis testing. That is, hypotheses are made about the source source location, which gives $\mathbf{G}(\omega)$. The likelihood are then measured. The hypothesis that results in the highest likelihood is determined to be the output of the SSL algorithm.

Instead of maximizing the likelihood in Eq. (20), we minimize the following negative log-likelihood:

$$J = \int_\omega J(\omega)d\omega. \quad (23)$$

Since we assume the probabilities over the frequencies are independent to each other, we may minimize each $J(\omega)$ separately by varying the unknown variable $S(\omega)$. Given $\mathbf{Q}^{-1}(\omega)$ is a Hermitian symmetric matrix, $\mathbf{Q}^{-1}(\omega) = \mathbf{Q}^{-H}(\omega)$, if we take derivative of $J(\omega)$ over $S(\omega)$, and set it to zero, we get:

$$S(\omega) = \frac{\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{X}(\omega)}{\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{G}(\omega)} \quad (24)$$

Insert the above $S(\omega)$ to $J(\omega)$, we get:

$$J(\omega) = J_1(\omega) - J_2(\omega) \quad (25)$$

$$J_1(\omega) = \mathbf{X}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{X}(\omega) \quad (26)$$

$$J_2(\omega) = \frac{[\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{X}(\omega)]^H\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{X}(\omega)}{\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{G}(\omega)} \quad (27)$$

Note $J_1(\omega)$ is unrelated to the hypothesized locations during hypothesis testing. Hence, the ML based SSL algorithm shall maximize:

$$J_2 = \int_\omega \frac{[\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{X}(\omega)]^H\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{X}(\omega)}{\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{G}(\omega)}d\omega \quad (28)$$

4. DISCUSSION

In order to compare the proposed algorithm with the existing approaches, let us first perform some simplifications on Eq. (28). Let us assume that the noises in the microphones are independent, thus $\mathbf{Q}(\omega)$ is a diagonal matrix:

$$\mathbf{Q}(\omega) = \text{diag}(\kappa_1,\cdots,\kappa_P), \quad (29)$$

with the i^{th} diagonal element as:

$$\kappa_i = \lambda_i + E\{|N_i(\omega)|^2\} = \gamma|X_i(\omega)|^2 + (1-\gamma)E\{|N_i(\omega)|^2\} \quad (30)$$

Eq. (28) can thus be written as:

$$J_2 = \int_\omega \frac{1}{\sum_{i=1}^P |\alpha_i(\omega)|^2/\kappa_i}\Big|\sum_{i=1}^P \frac{\alpha_i^*(\omega)}{\kappa_i}X_i(\omega)e^{j\omega\tau_i}\Big|^2 d\omega \quad (31)$$

Fig. 1. RoundTable device and its captured images. (a) The RoundTable device. (b) The captured panoramic images.

The gain factor $\alpha_i(\omega)$ can be accurately measured in some applications. For applications where it is unknown, we may assume it as a positive real number and estimate it as follows:

$$|\alpha_i(\omega)|^2|S(\omega)|^2 \approx |X_i(\omega)|^2 - \kappa_i, \quad (32)$$

where both sides represent the power of the signal received at microphone i without the combined noise (noise and reverberation). Therefore,

$$\alpha_i(\omega) = \sqrt{(1-\gamma)(|X_i(\omega)|^2 - E\{|N_i(\omega)|^2\})}/|S(\omega)|, \quad (33)$$

Insert Eq. (33) to Eq. (31), we get:

$$J_2 = \int_\omega \frac{\Big|\sum_{i=1}^P \frac{1}{\kappa_i}\sqrt{|X_i(\omega)|^2 - E\{|N_i(\omega)|^2\}}X_i(\omega)e^{j\omega\tau_i}\Big|^2}{\sum_{i=1}^P \frac{1}{\kappa_i}|X_i(\omega)|^2 - E\{|N_i(\omega)|^2}d\omega \quad (34)$$

In the cases that the signal to noise ratio (SNR) is very high, we have $|X_i(\omega)|^2 \gg E\{|N_i(\omega)|^2\}$. It is easy to verify that Eq. (34) can be simplified to the SRP-PHAT algorithm, Eq. (8).

The connection between the proposed ML algorithm and the ML algorithm in Eq. (9) is not straightforward. Recall in their original derivation, Brandstein et al. [1] gave the variance of the estimated phase for a particular frequency as:

$$Var[\theta_i(\omega)] = E\{|N_i(\omega)|^2\}/|X_i(\omega)|^2. \quad (35)$$

If we ignore reverberation, i.e., set $\gamma = 0$, and assume noise is relatively small compared with the signal (the same assumptions were made in [1]), Eq. (34) can be written as:

$$J_2 = \int_\omega \frac{\Big|\sum_{i=1}^P \frac{e^{\theta_i(\omega)}e^{j\omega\tau_i}}{E\{|N_i(\omega)|^2\}/|X_i(\omega)|^2}\Big|^2}{\sum_{i=1}^P |X_i(\omega)|^2/E\{|N_i(\omega)|^2\}}d\omega. \quad (36)$$

Therefore, the phase term of each microphone $e^{\theta_i(\omega)}e^{j\omega\tau_i}$ is indeed weighted by the inverse of the phase variance (Eq. (35)). Hence the previous ML algorithm is conceptually similar to the proposed algorithm. On the other hand, the proposed algorithm differs from the previous method in the additional frequency-dependent weighting (denominator in Eq. (36)). It also has a more rigorous derivation and is a true ML algorithm for multiple microphones.

The ML SSL framework presented in [13] is closely related to ours. There the goal is to estimate not only the sound source location, but also its directionality. They used a similar model as Eq. (12), but without the reverberation term. The noise covariance matrix is assumed as diagonal, $\mathbf{Q}(\omega) = \sigma\mathbf{I}$, where σ is independent of the microphone index and frequency, which led to a simplified target function as:

$$J_2 = \int_\omega |\sum_{i=1}^P \alpha^*(\omega)X_i(\omega)e^{j\omega\tau_i}|^2 d\omega. \quad (37)$$

It is not difficult to verify that with all their assumptions, Eq. (37) can be easily obtained from Eq. (31).

SRP-PHAT		Alg. in [18]		$\gamma = 0.2$				$\gamma = 0.5$			
				SRP-RUI		ML		SRP-RUI		ML	
<6°	<14°	<6°	<14°	<6°	<14°	<6°	<14°	<6°	<14°	<6°	<14°
81.73%	88.13%	80.55%	86.85%	83.06%	89.76%	**83.49%**	**90.13%**	82.76%	89.31%	83.03%	89.96%

Fig. 2. Experimental results of SSL accuracy on the real-world dataset. Cells with bold fonts indicate best performance in the group.

5. EXPERIMENTAL RESULTS

We test the performance of the proposed SSL algorithm, in particular, Eq. (34), on both synthetic and real-world datasets. Due to page limits, we report the results on real-world dataset only, and refer the reader to [19] for more detailed results.

The two benchmark algorithms we use to compare with the proposed method are SRP-PHAT (Eq. (8)) and its improved version from [18] (Eq. (11)). Hereafter the second benchmark algorithm is referred as SRP-RUI. Note SRP-PHAT is a special case of SRP-RUI when $\gamma = 1.0$, while SRP-RUI is a special case of the proposed ML-based SSL algorithm when $\alpha_i(\omega) \equiv \alpha(\omega), i = 1, \cdots, P$, and the frequency weightings are ignored.

We test the three SSL algorithms on 99 real-world meetings captured by the RoundTable device. Fig. 1 shows the device as well as two example panoramic images of the meeting rooms. SSL is used in RoundTable to help frame the speaker in a high-resolution video output. One difficulty of SSL for the RoundTable device is the directional microphones deployed to capture better audio. For microphones facing away from the speaker, the phase may not be very reliable. In [18], the authors combated the issue by selecting a subset of the microphones for SSL. In this paper, we will still use all the microphones, since the ML-based SSL should have weighted different microphones based on their SNR automatically. We will compare our results with [18].

The meetings are 4 minutes each, captured in about 50 different meeting rooms in order to test the robustness of the SSL algorithms in different environments. The noise levels of the rooms and the distances from the speakers to the devices vary significantly, causing the input SNR to range from 5 dB to 25 dB. The speaker locations of 6706 audio frames are labeled manually based on the corresponding face locations in the panoramic image. We report the results on the percentage of frames that are within 6° and 14° of the ground truth azimuth angle. This is good enough for the purpose of speaker pointing in RoundTable.

The experimental results are shown in Fig. 2. It can be seen that the proposed ML-based SSL performs the best on this challenging dataset. The improvement of ML-based SSL over SRP-PHAT is about 2%. After examining the sequences, we found for many high SNR sequences the two algorithms perform almost the same. However, ML-based SSL achieves significant improvement on those noisy sequences. The algorithm in [18] is more efficient than SRP-PHAT, however the performance is slightly worse. This may be due to the limited data available at that time.

6. CONCLUSION

We have shown a ML-based SSL algorithm that is efficient to compute (Eq. (34)) yet works very well in practice. One future work we are working on is to extend the current framework to multi-source scenarios, which happens surprisingly often during daily meetings.

7. REFERENCES

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

[2] W. Wahlster, N. Reithinger, and A. Blocher, "Smartkom: multimodal communication with a life-like character," in *Proc. Eurospeech*, 2001.

[3] S. Basu, B. Clarkson, and A. Pentland, "Smart headphones: enhancing auditory awareness through robust speech detection and source localization," in *Proc. of IEEE ICASSP*, 2001.

[4] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. of IEEE ICASSP*, 1997.

[5] R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L.W. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverbert, "Distributed meetings: a meeting capture and broadcasting system," in *Proc. ACM Conf. on Multimedia*, 2002.

[6] J. Weng and K. Y. Guentchev, "Three-dimensional sound localization from a compact non-coplanar array of microphones using tree-based learning," *The Journal of the Acoustical Society of America*, vol. 110, no. 1, pp. 310–323, 2001.

[7] J. Kleban, "Combined acoustic and visual processing for video conferencing systems," Tech. Rep., The State University of New Jersy, Rutgers, 2000.

[8] P. Georgiou, C. Kyriakakis, and P. Tsakalides, "Robust time delay estimation for sound source localization in noisy environments," in *Proc. of WASPAA*, 1997.

[9] T. Gustafsson, B. Rao, and M. Trivedi, "Source localization in reverberant environments: performance bounds and ml estimation," in *Proc. of ICASSP*, 2001.

[10] D. Li and S. Levinson, "Adaptive sound source localization by two microphones," in *Proc. of Int. Conf. on Robotics and Automation*, 2002.

[11] I. Ziskind and M. Wax, "Maximum likelihood localization of multiple sources by alternating projection," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 36, no. 10, pp. 1553–1560, 1988.

[12] K. Harmanci, J. Tabrikian, and J. L. Krolik, "Relationships between adaptive minimum variance beamforming and opitmal source localization," *IEEE Trans. on Signal Processing*, vol. 48, no. 1, pp. 1–12, 2000.

[13] B. Mungamuru and P. Aarabi, "Enhanced sound localization," *IEEE Trans. on Systems, Man and Cybernetics – Part B: Cybernetics*, vol. 34, no. 13, pp. 1526–1540, 2004.

[14] X. Sheng and Y.-H. Hu, "Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks," *IEEE Trans. on Signal Processing*, vol. 53, no. 1, pp. 44–53, 2005.

[15] M. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. of ICASSP*, 1997.

[16] Y. Rui and D. Florêncio, "Time delay estimation in the presence of correlated noise and reverberation," in *Proc. of ICASSP*, 2004.

[17] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-24, no. 4, pp. 320–327, 1976.

[18] Y. Rui, D. Florêncio, W. Lam, and J. Su, "Sound source localization for circular arrays of directional microphones," in *Proc. of ICASSP*, 2005.

[19] C. Zhang, Z. Zhang, and D. Florêncio, "A maximum likelihood framework for multi-microphone sound source localization," Tech. Rep., Microsoft Research, 2006 (to appear).