Microphone-array localization error estimation with application to sensor placement

Michael S. Brandstein, a) John E. Adcock, b) and Harvey F. Silverman c)

Laboratory for Engineering Man/Machine Systems, Division of Engineering, Brown University, Providence, Rhode Island 02912

(Received 30 June 1995; revised 4 January 1996; accepted 22 January 1996)

This paper presents a means for predicting the error region associated with a speech-source location estimate obtained from a set of microphones in a room environment. The error predictor presented is derived assuming a specific source-sensor geometry consisting of pairs of closely spaced sensors for which a delay estimate associated with the potential source has been evaluated. The accuracy of the predictor is evaluated through a set of Monte Carlo simulations and an application of the predictor to microphone-array design in the context of a video-conferencing scenario is presented. © 1996 Acoustical Society of America.

PACS numbers: 43.72.Kb, 43.38.Hz

INTRODUCTION

A steerable array of microphones has the potential to replace the traditional head-mounted or desk-stand microphone as the input transducer system for acquiring speech data in many applications. An array of microphones has a number of advantages over a single-microphone system. First, it may be electronically aimed to provide a high-quality signal from a desired source location while it simultaneously attenuates interfering talkers and ambient noise. In this regard, an array has the potential to outperform a single, well-aimed, highly directional microphone. Second, an array system does not necessitate local placement of transducers, will not encumber the talker with a hand-held or head-mounted microphone, and does not require physical movement to alter its direction of reception. These features make it advantageous in settings involving multiple or moving sources. Finally, it has potential capabilities that a single microphone does not; namely automatic detection, location, and tracking of active talkers in its receptive area. Existing array systems have been used in a number of applications. 1,2 These include teleconferencing, 3–6 speech recognition, 7–10 speaker identification, 11 speech acquisition in an automobile environment, 12,13 sound capture in reverberant enclosures, 14–16 large-room recording-conferencing, 17 acoustic surveillance, 18,19 and hearing aid devices. 20 These systems also have the potential to be beneficial in several other environments, the performing arts and sporting communities, for instance.

An essential requirement of these sensor array systems is the ability to locate and track a speech source and accordingly various methods for talker localization in a room environment have been investigated. 16,18,21–23 For audio-based applications, an accurate fix on the primary talker, as well as knowledge of any interfering talkers or coherent noise sources, is necessary to effectively steer the array, enhancing a given source while simultaneously attenuating those deemed undesirable. Source-location information may be used to discriminate individual speakers in a multisource 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 is the use of speaker location estimates for aiming one or more cameras in a video-conferencing system. In this application, the automated localization system eliminates the need for human camera operators.

Given a source-location estimate, a measure of the region of spatial uncertainty related to the estimate is essential before the information can be judiciously employed in a practical application. For instance, given a source-location estimate, knowledge of the associated error region is required for designing the array beam pattern; allowing for a sufficiently wide beam to avoid audio dropout due to misaim while simultaneously restricting the beam size to provide maximum attenuation of interfering sources. Similarly, in pointing a video camera, this error information is useful for selecting camera magnification, permitting close-up views of a talker without overzooming. This data may also be used to aid in array design. As will be demonstrated in Sec. IV, the estimate uncertainty may be incorporated into a process for determining optimal microphone placement.

The assessment of array localization accuracy has been studied extensively in the context of sonar and radar applications. 24–29 These analyses concentrate primarily on linear sensor arrays, evaluating location estimates with regard to range and bearing error. This information is also of interest for navigational systems, 30–33 where a location estimate is calculated from clocking signals transmitted from known positions. These applications vary in key respects from the speech-source/room-environment scenario addressed below.

This paper presents a means for predicting the error region associated with a speech-source location estimate obtained from a set of microphones placed within a room environment. The error predictor presented is derived assuming the specific source-sensor geometry utilized in Ref. 21; the basic unit of which consists of a pair of closely spaced sen-
sensors and a single delay estimate associated with the potential source. Delay estimates are evaluated with respect to the particular sensor pair. Unlike the case of navigational systems, there is no attempt made to define time-difference of arrival (TDOA) values relative to a single reference sensor or an absolute scale. This philosophy is motivated by several arguments. Primarily, in a near-field source environment, source directionality can create significant signal dissimilarities at spatially distant sensors. In the interest of obtaining accurate and reliable TDOA estimates, the individual sensors in each pair must be kept close together. Additionally, the precision of the location estimate is dependent upon the placement of the sensors relative to the actual source location and the acoustic environment. This may require placing sensors in a wide variety of positions throughout the enclosure. In general, sensor geometries will not be limited to linear arrays, preventing an analysis of location estimate accuracy in terms of range and bearing alone. Given only a fixed number of sensors arranged in spatially local sensor pairs, it is not robust and frequently not possible, to evaluate all the TDOA figures relative to a single sensor location. The proposed sensor-pair geometry addresses the problem of source localization given these autonomous sensor pair-TDOA units.

The following section describes the source-localization problem addressed by this work and presents a statistically optimal location estimator. Section II contains the derivation of an error region predictor which is then evaluated through a set of simulations in Sec. III. Section IV illustrates the specific application of the predictor in the context of array design in a video-conferencing scenario. Finally, Sec. V presents some conclusions and suggestions for future work.

I. SOURCE LOCALIZATION PROBLEM

The locationing problem addressed here may be stated as follows. There are \( N \) pairs of sensors \( m_{i1} \) and \( m_{i2} \) for \( i = 1, \ldots, N \). The ordered triplet \((x, y, z)\) of spatial coordinates for the sensors will be denoted by \( m_{i1} \) and \( m_{i2} \), respectively. For each sensor pair, a TDOA estimate, \( \tau_i \), for a signal source located at \( s \) is available. The true TDOA associated with a source, \( s \), and the \( i \)th sensor pair is given by

\[
T_i = \frac{|s - m_{i1}| - |s - m_{i2}|}{c},
\]

where \( c \) is the speed of propagation in the medium. In practice, \( \tau_i \) represents a corrupted version of the true TDOA and in general, \( \tau_i \neq T_i \). In addition to the \( \tau_i \), a variance estimate, \( \sigma^2_i \), associated with each TDOA is also assumed to be available as a byproduct of the time-delay estimation procedure. This variance estimate is generally a function of the signal content and sensor SNR conditions as well as the source radiation pattern and the physical conditions of the room. Throughout this discussion \( \tau_i \) is assumed to be an unbiased estimate of the true TDOA. This assumption is reasonable under a broad range of physical conditions. For acoustically live environments reverberation and multipath affect TDOA estimation in a manner similar to that of uncorrelated white noise under diminished SNR situations.

These conditions would then be incorporated into an inflated TDOA variance estimate.

Given these \( N \) sensor pair, TDOA-estimate combinations:

\[
\{m_{i1}, m_{i2}\}, \tau_i, \sigma^2_i, \quad \text{for } i = 1, \ldots, N,
\]

it is desired to obtain an estimate of the source location, \( \hat{s} \). If the TDOA estimates are assumed to be independently corrupted by additive zero-mean Gaussian noise, the maximum likelihood (ML) estimate \( \hat{s}_\text{ML} \) is found through minimization of a least-squares error criterion\(^{15}\) denoted here by \( J_{\text{ML}}(\hat{s}) \):

\[
J_{\text{ML}}(\hat{s}) = \sum_{i=1}^{N} \frac{1}{\sigma^2_i} \cdot |\tau_i - T_i(\{m_{i1}, m_{i2}\}, \hat{s})|^2.
\]

While other localization error criteria are amenable to this localization problem and have been shown to yield advantageous results under the least favorable conditions,\(^{21}\) only the ML error criterion will be considered here.

II. ESTIMATION OF LOCALIZATION ERROR REGION

In what follows it is assumed that the true source location \( s \) is known and that the goal is to develop a statistical analysis of the precision associated with the source-location estimate \( \hat{s} \). First, the displacement associated with the source estimate is expressed in terms of the true location and the array geometry, and then analyzed relative to the localization-error criterion.

A. Displacement geometry

For a pair of sensors, \( m_{i1} \) and \( m_{i2} \), the midpoint \( m_i \) and unit axis \( a_i \) are given by

\[
m_i = \frac{m_{i1} + m_{i2}}{2}, \quad a_i = \frac{m_{i1} - m_{i2}}{|m_{i1} - m_{i2}|}.
\]

Figure 1 depicts the relationship between the true source location \( s \) and an estimate of the location \( \hat{s} \) relative to this \( i \)th sensor pair. \( R_i \) is defined as the distance from \( s \) to \( m_i \) and \( \theta_i \) as the angle between the directed line segment \( s - m_i \) and the sensor pair axis \( a_i \). The values \( R_i \) and \( \theta_i \) are defined similarly for the location estimate \( \hat{s} \). The three-dimensional Cartesian displacement vector from \( s \) to \( \hat{s} \) is denoted by

\[
\Delta s = \begin{bmatrix}
\Delta s_x \\
\Delta s_y \\
\Delta s_z
\end{bmatrix}.
\]

For sources with a large source-range to sensor-separation ratio \( (R_i/|m_{i1} - m_{i2}|) \gg 1 \), the direction angles, \( \theta_i \) and \( \theta_i \), are well approximated by

\[
\hat{\theta}_i = \cos^{-1} \left( \frac{c \cdot T_i(\{m_{i1}, m_{i2}\}, \hat{s})}{|m_{i1} - m_{i2}|} \right),
\]

\[
\theta_i = \cos^{-1} \left( \frac{c \cdot T_i(\{m_{i1}, m_{i2}\}, s)}{|m_{i1} - m_{i2}|} \right),
\]

These conditions would then be incorporated into an inflated TDOA variance estimate.

Given these \( N \) sensor pair, TDOA-estimate combinations:

\[
\{m_{i1}, m_{i2}\}, \tau_i, \sigma^2_i, \quad \text{for } i = 1, \ldots, N,
\]

it is desired to obtain an estimate of the source location, \( \hat{s} \). If the TDOA estimates are assumed to be independently corrupted by additive zero-mean Gaussian noise, the maximum likelihood (ML) estimate \( \hat{s}_\text{ML} \) is found through minimization of a least-squares error criterion\(^{15}\) denoted here by \( J_{\text{ML}}(\hat{s}) \):

\[
J_{\text{ML}}(\hat{s}) = \sum_{i=1}^{N} \frac{1}{\sigma^2_i} \cdot |\tau_i - T_i(\{m_{i1}, m_{i2}\}, \hat{s})|^2.
\]

While other localization error criteria are amenable to this localization problem and have been shown to yield advantageous results under the least favorable conditions,\(^{21}\) only the ML error criterion will be considered here.

II. ESTIMATION OF LOCALIZATION ERROR REGION

In what follows it is assumed that the true source location \( s \) is known and that the goal is to develop a statistical analysis of the precision associated with the source-location estimate \( \hat{s} \). First, the displacement associated with the source estimate is expressed in terms of the true location and the array geometry, and then analyzed relative to the localization-error criterion.

A. Displacement geometry

For a pair of sensors, \( m_{i1} \) and \( m_{i2} \), the midpoint \( m_i \) and unit axis \( a_i \) are given by

\[
m_i = \frac{m_{i1} + m_{i2}}{2}, \quad a_i = \frac{m_{i1} - m_{i2}}{|m_{i1} - m_{i2}|}.
\]

Figure 1 depicts the relationship between the true source location \( s \) and an estimate of the location \( \hat{s} \) relative to this \( i \)th sensor pair. \( R_i \) is defined as the distance from \( s \) to \( m_i \) and \( \theta_i \) as the angle between the directed line segment \( s - m_i \) and the sensor pair axis \( a_i \). The values \( R_i \) and \( \theta_i \) are defined similarly for the location estimate \( \hat{s} \). The three-dimensional Cartesian displacement vector from \( s \) to \( \hat{s} \) is denoted by

\[
\Delta s = \begin{bmatrix}
\Delta s_x \\
\Delta s_y \\
\Delta s_z
\end{bmatrix}.
\]

For sources with a large source-range to sensor-separation ratio \( (R_i/|m_{i1} - m_{i2}|) \gg 1 \), the direction angles, \( \theta_i \) and \( \theta_i \), are well approximated by

\[
\hat{\theta}_i = \cos^{-1} \left( \frac{c \cdot T_i(\{m_{i1}, m_{i2}\}, \hat{s})}{|m_{i1} - m_{i2}|} \right),
\]

\[
\theta_i = \cos^{-1} \left( \frac{c \cdot T_i(\{m_{i1}, m_{i2}\}, s)}{|m_{i1} - m_{i2}|} \right),
\]
FIG. 1. The relationship between the true source location \( \mathbf{s} \) and an estimate of the location \( \mathbf{s}^\prime \) relative to the \( i \)th sensor pair.

and \( \hat{\theta}_i \) is related to the positional vectors via the dot product:

\[
\hat{R}_i \cos \hat{\theta}_i = (\mathbf{s} + \mathbf{\Delta s} - \mathbf{m}_i) \cdot \mathbf{a}_i = (\mathbf{s} - \mathbf{m}_i) \cdot \mathbf{a}_i + \mathbf{\Delta s} \cdot \mathbf{a}_i = R_i \cos \theta_i + \mathbf{\Delta s} \cdot \mathbf{a}_i. \tag{6}
\]

Following Ref. 32, \( \hat{R}_i \) is represented by its first-order Taylor series expansion about the true source location:

\[
\hat{R}_i \approx R_i + \left( \frac{\mathbf{s} - \mathbf{m}_i}{R_i} \right) \cdot \mathbf{\Delta s}. \tag{7}
\]

This linearization of the source range requires that the source-location estimate be sufficiently close to the true location (i.e., small \( \mathbf{\Delta s} \)) such that the error introduced in this approximation is negligible in comparison to localization errors due to TDOA estimate errors. As will be shown, this requirement is met in practice. Additionally, \( \mathbf{\hat{s}} \) is an inherently biased estimate of the true location. By truncating the series expansion the nonlinear terms responsible for this effect are eliminated, and accordingly, the approximation that follows does not account for any biasing found in the actual location estimate.

Substituting (7) into (6) yields

\[
R_i \cos \hat{\theta}_i + \left( \frac{\mathbf{s} - \mathbf{m}_i}{R_i} \right) \cdot \mathbf{\Delta s} \cos \theta_i = R_i \cos \theta_i + \mathbf{\Delta s} \cdot \mathbf{a}_i,
\]

or equivalently

\[
\cos \hat{\theta}_i - \cos \theta_i = \left[ \frac{\mathbf{a}_i}{R_i} - \left( \frac{\mathbf{s} - \mathbf{m}_i}{R_i^2} \right) \cdot \mathbf{\Delta s} \right]. \tag{8}
\]

From the assumption of a small \( \mathbf{\Delta s} \), \( \cos \theta_i / R_i^2 \approx \cos \hat{\theta}_i / R_i^2 \). Applying this and (5) to the cosine terms on the left side of (8) produces

\[
\frac{1}{|\mathbf{m}_2 - \mathbf{m}_1|} \left[ T(\{\mathbf{m}_1, \mathbf{m}_2\}, \mathbf{s}) - T(\{\mathbf{m}_1, \mathbf{m}_2\}, \mathbf{s}) \right] = \left[ \frac{\mathbf{a}_i}{R_i} - \left( \frac{\mathbf{s} - \mathbf{m}_i}{R_i^2} \right) \cdot \mathbf{\Delta s} \right].
\]

And finally

\[
\begin{aligned}
\left[ T(\{\mathbf{m}_1, \mathbf{m}_2\}, \mathbf{s}) - T(\{\mathbf{m}_1, \mathbf{m}_2\}, \mathbf{s}) \right] &= \left( \frac{\mathbf{a}_i}{R_i} - \left( \frac{\mathbf{s} - \mathbf{m}_i}{R_i^2} \right) \cdot \mathbf{\Delta s} \right) \\
 &= \mathbf{h}'_i \cdot \mathbf{\Delta s},
\end{aligned} \tag{9}
\]

where \( \mathbf{h}'_i \) is the \((1 \times 3)\) vector relating the difference in TDOA for the \( i \)th sensor pair to the estimate displacement vector.

It will be useful to express (9) for the \( N \) sensor pairs via matrix notation. The \((N \times 1)\) vector of TDOA differences is denoted by \( \mathbf{\Delta \tau} \) and the \((N \times 3)\) matrix composed of the \( \mathbf{h}'_i \) vectors will be given by \( \mathbf{H} \), i.e.,

\[
\mathbf{\Delta \tau} = \mathbf{H} \mathbf{\Delta s}, \tag{10}
\]

B. Source estimate based upon the ML criterion

The case where the source location is estimated by minimization of the ML error criterion given in (2) is now examined; \( \mathbf{\hat{s}} = \mathbf{\hat{s}}_{\text{ML}} \).

The ML error criterion (3) may be rewritten as

\[
\mathbf{J}_{\text{ML}}(\mathbf{\hat{s}}) = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \cdot [T(\{\mathbf{m}_1, \mathbf{m}_2\}, \mathbf{s})]^2
\]

\[
= \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \cdot [T(\{\mathbf{m}_1, \mathbf{m}_2\}, \mathbf{s})^2 + T(\{\mathbf{m}_1, \mathbf{m}_2\}, \mathbf{s})^2] \tag{11}
\]

Defining \( \mathbf{W} \) as the \((N \times N)\) diagonal matrix of the reciprocal TDOA estimate variances and \( \mathbf{\Delta \tau} \) to be the \((N \times 1)\) vector of differences between the estimated TDOA and true TDOA,

\[
\mathbf{\Delta \tau} = \begin{bmatrix} T(\{\mathbf{m}_1, \mathbf{m}_2\}, \mathbf{s}) \\ T(\{\mathbf{m}_2, \mathbf{m}_3\}, \mathbf{s}) \\ \vdots \\ T(\{\mathbf{m}_{N-1}, \mathbf{m}_N\}, \mathbf{s}) \end{bmatrix}
\]

And finally

\[
\begin{aligned}
\mathbf{J}_{\text{ML}}(\mathbf{\hat{s}}) &= \frac{1}{\sigma_i^2} \cdot [T(\{\mathbf{m}_1, \mathbf{m}_2\}, \mathbf{s})^2 + T(\{\mathbf{m}_1, \mathbf{m}_2\}, \mathbf{s})^2] \\
&= \frac{\mathbf{W}}{\sigma_i^2} \cdot (\mathbf{\Delta \tau}^2)
\end{aligned}
\]

where \( \mathbf{W} \) is the diagonal matrix of the reciprocal TDOA estimate variances.
Equation (11) is rewritten as

$$J_{ML}(\hat{s}) = ((\Delta \tau - \Delta \tau_0)'W(\Delta \tau - \Delta \tau_0)).$$

The ML criterion is assumed to be minimized when $\hat{s} = \hat{s}_{ML}$, which gives

$$J_{ML}(\hat{s}) = \min(J_{ML}(\hat{s}))$$

$$= ((\Delta \tau - \Delta \tau_0)'W(\Delta \tau - \Delta \tau_0))$$

$$= ((\Delta \tau - H\Delta s_{ML})'W(\Delta \tau - H\Delta s_{ML}).$$

The right side of this equation is identical in form to the weighted-linear-least-squares-error and can be shown to be minimized when

$$\Delta s_{ML} = (H'WH)^{-1}H'WB\Delta \tau_r.$$

Therefore, minimization of $J_{ML}$ would result in a $\Delta s_{ML}$ as given above. Equation (13) relates the displacement vector associated with a location estimate $\hat{s}_{ML}$ to the TDOA estimates that would produce this particular estimate via minimization of the nonlinear LS error criterion $J_{ML}$.

The covariance of $\Delta s_{ML}$ is given by

$$\text{cov}\{\Delta s_{ML}\} = E[(\Delta s_{ML} - E(\Delta s_{ML}))(\Delta s_{ML} - E(\Delta s_{ML})).]$$

The delay estimates have been assumed to be corrupted by a zero-mean, uncorrelated noise source and therefore $E(\Delta \tau_r) = 0$. Substituting this and (13) into (14) yields

$$\text{cov}\{\Delta s_{ML}\} = E[\Delta s_{ML}\Delta s_{ML}]$$

$$= (H'WH)^{-1}H'WE(\Delta \tau, \Delta \tau_r)W'H((H'WH)^{-1})'$$

$$= (H'WH)^{-1}H'WE(\Delta \tau, \Delta \tau_r)W'H(H'WH)^{-1},$$

because $E(\Delta \tau_r) = 0$, $\text{cov}\{\Delta \tau_r\} = E(\Delta \tau, \Delta \tau_r)$ which is an $(N \times N)$ diagonal matrix with diagonal elements $\sigma^2_\tau$. Similarly, the weighting coefficients that comprise the diagonal elements of $W$ were selected to be $(1/\sigma^2_\tau)$, and thus $E(\Delta \tau, \Delta \tau_r)W' = I_N$. Equation (15) therefore simplifies to

$$\text{cov}\{\Delta s_{ML}\} = (H'WH)^{-1}(H'WH)(H'WH)^{-1}$$

$$= (H'WH)^{-1}.$$  

Equation (16) predicts the covariance of the $\hat{s}_{ML}$ estimate given knowledge of the source and sensor locations as well the TDOA estimate variances.

III. ANALYSIS OF ESTIMATE ERROR PREDICTORS

To evaluate the accuracy of (16) as a predictor of the estimator’s true covariance, two sets of simulations were conducted with the varying parameter being the positions of the sensor pairs.

A. Evaluation 1

In this first computer simulation a ten-element bi-linear sensor array with 0.5-m spacings was used. The array was centered on one wall in a 6-m x 6-m x 4-m idealized, anechoic rectangular room as depicted in Fig. 2(a).

The eight pairs of diagonally adjacent sensors were selected as the sensor pairs. Monte Carlo simulations were performed; 100 independent location estimates were computed for each of 36 source locations within the room. Source locations were spaced one meter apart in two horizontal planes at 2 m and 3 m high. For each of the 100 trials at each source location the true TDOA values for each sensor pair were calculated and then corrupted by uncorrelated additive white noise.
Gaussian noise. The corrupting noise level at each sensor pair was fixed at a moderate level, a standard deviation of 10^{-2} m when scaled by \( c \). The ML location estimate \( \hat{s}_{\text{ML}} \) was then calculated for each trial via a quasi-Newton algorithm constrained to search within the physical dimensions of the room.

Figure 3 displays the results of these simulations. Each plot in this figure is a view from above the room. The sensor array is indicated by the five circles on the left vertical axis. In Fig. 3(a), the 3600 (36 locations, 100 estimates per location) \( \hat{s}_{\text{ML}} \) estimates have been plotted with dots. Because the room is symmetric about the midpoint of the array, only one height was used for the locations in each half-plane. In each of these plots, the lower horizontal half-plane contains source locations 2 m high, level with the midline of the sensor array. The sources in the upper half-plane are 3 m high, 1 m above the midline of the sensor array.

Figure 3(b) shows the principal component vectors of the predicted covariance matrix scaled to 2.5 standard deviations with thick lines, and the principal component vectors of the distribution of measured source locations. The principal components of the measured distributions are offset vertically by a small amount for graphical clarity. For each source location, the predicted error covariance matrix was calculated via (16). An eigenvector–eigenvalue decomposition of the \( (3 \times 3) \) matrix yields its principal components vectors.\(^{36}\)

Geometrically, if \( \text{cov} \{ \Delta s_{\text{ML}} \} \) is positive definite with eigenvalue-eigenvector pairs \( (\lambda_i, e_i) \) for \( i = 1, 2, 3 \), all the \( (3 \times 1) \) vectors \( x \) which satisfy

\[
(x - \overline{x})^T (\text{cov} \{ \Delta s_{\text{ML}} \})^{-1} (x - \overline{x}) = h^2
\]

define a hyperellipsoid centered about \( \overline{x} \) with axes \( \pm h \sqrt{\lambda_i} e_i \). The eigenvalues correspond to the variance of the data set projected onto the corresponding eigenvector or principal component. Setting \( h = 2.5 \) in the above expression will therefore generate a hyperellipsoid with axes extending 2.5 standard deviations in either direction from the center of the conic along each of the principal component vectors.

<table>
<thead>
<tr>
<th>Point label</th>
<th>Standard deviations of principal components (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Predicted</td>
</tr>
<tr>
<td>A</td>
<td>1st</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
</tr>
<tr>
<td></td>
<td>3rd</td>
</tr>
<tr>
<td></td>
<td>total</td>
</tr>
<tr>
<td>B</td>
<td>1st</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
</tr>
<tr>
<td></td>
<td>3rd</td>
</tr>
<tr>
<td></td>
<td>total</td>
</tr>
<tr>
<td>C</td>
<td>1st</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
</tr>
<tr>
<td></td>
<td>3rd</td>
</tr>
<tr>
<td></td>
<td>total</td>
</tr>
<tr>
<td>D</td>
<td>1st</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
</tr>
<tr>
<td></td>
<td>3rd</td>
</tr>
<tr>
<td></td>
<td>total</td>
</tr>
<tr>
<td>E</td>
<td>1st</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
</tr>
<tr>
<td></td>
<td>3rd</td>
</tr>
<tr>
<td></td>
<td>total</td>
</tr>
<tr>
<td>F</td>
<td>1st</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
</tr>
<tr>
<td></td>
<td>3rd</td>
</tr>
<tr>
<td></td>
<td>total</td>
</tr>
</tbody>
</table>
the distribution of source estimates possesses a trivariate normal density, a given estimate would have a 0.9 probability of falling on or within such a hyperellipsoid. In each case, the estimator has been assumed to be zero biased and thus the center of the hyperellipsoid is the given source location. The lines in Fig. 3(b) display the scaled principal component vectors which correspond to the axes of the hyperellipsoid associated with the predicted error covariance of each source location shown in Fig. 3(a).

Table I presents detailed numerical data for selected source locations in Fig. 3. In each case, the principal component standard deviations are listed for the predicted error covariance as well as for the sample covariances associated with the set of $\hat{s}_{ML}$ estimates. The fourth row gives the total standard deviation. This value is calculated as the square root of the component variance summation and is equivalent to the square root of the trace of the covariance matrix.

Several characteristics are apparent from Fig. 3 and Table I. The predicted error covariance is generally within a few centimeters of the observed covariance of the $\hat{s}_{ML}$ estimator. As the figure suggests and the table quantities, disparities between the predicted and observed are most extreme in those cases involving relatively large error variances, where the linearity assumptions used in the derivation of the error covariance predictor are less valid. Consider the source point labeled “D,” located at (3.5 m, 0.5 m) and at a height of 2 m. The total observed standard deviation is 33.3 cm for the estimator while the predicted total is 35.7 cm. The 2.4-cm difference is largely accounted for by the first principle component. At the other extreme is the point “B” at (1.5 m, 1.5 m) and height 2 m. The disparity between predicted and observed total standard deviation is only 0.1 cm. For sources positioned near the boundary of the room, this disparity between observation and prediction may be due, in part, to artificially low observed covariance values brought about by the search constraint placed on the ML error criterion minimizer. Those source locations that are estimated to be outside of the physical room are placed at the room boundary and the spread of source estimates is subsequently skewed. This effect is apparent in the statistics for point “F” where there is a visible cluster of estimate points at the wall of the room and the observed first-principal component standard deviations are sizably less than the predicted value.

As expected, the source-estimation procedure is most accurate for sources close to and within the dimensions of the sensor array. The precision of the location estimates is extremely sensitive to the bearing for sources close to endfire, and the accuracy of the range estimates drops rapidly as the true source range increases. These observations are consistent with results reported for standard linear arrays, as in Ref. 37. For a fixed $(x, y)$ position relative to the floor, the variation in height of the half-planes had little effect on the estimator’s precision. An exception to this rule being the source location labeled “A” at location (0.5 m, 5.5 m) and height 3 m and its symmetric counterpart at (0.5 m, 0.5 m) and height 2 m. While the former is farther from the sensors than the latter, it possesses a milder bearing condition relative to the array. As the figure illustrates, this small improvement in bearing angle has a dramatic effect on the error spread of the source’s location estimates in comparison to its counterpart’s.

B. Evaluation 2

In the previous simulation, the source locations that were estimated most precisely possessed two key features: they were broadside (or nearly broadside) to the sensor array, and they were not very distant from the sensors. Short of placing physical obstacles at positions considered undesirable, the set of potential source locations within a room cannot be dictated. However, there may be a great deal of freedom in the placement of the sensors. The results of evaluation 1 motivated the choice of the array configuration illustrated by Fig. 2(b) in which a 0.5-m×0.5-m square array has been centered along each wall of the same 6-m×6-m×4-m anechoic, rectangular room. This sensor arrangement provides for an improved coverage of the room environment. The vast majority of potential source locations are close and at a broadside angle to at least one sensor pair. The diagonal combinations within each subarray were used as the sensor pairs, yielding the same number of TDOA estimates (eight) and the same sensor spacings (0.5√2 m) as in the previous experiment. Monte Carlo simulations were performed in a manner identical to those of evaluation 1. Now that the array configuration is symmetric in two dimensions, the source locations in each of the four parallel quarter planes are at a single height, ranging from the midpoint of the room, 2 m, to a half-meter short of the ceiling, 3.5 m.

The results of this experiment are presented in Fig. 4 and Table II. Once again, the expression (16) accurately predicts the results of the Monte Carlo simulations. Discrepancies between observed and predicted values continue to be greatest in the large variance cases. Source height has little effect on the overall estimation precision, except in the cases where altering the source height significantly alters the source’s bearing angle relative to a sensor pair. In short, the trends from evaluation 1 are also seen with this alternative sensor arrangement. However, the overall source-localization error has been reduced significantly as a result of the more judicious placement of sensors; the predicted mean total standard deviation decreased from 30.4 cm in evaluation 1 to 7.3 cm in evaluation 2. Also the hyperellipsoids in evaluation 2 are considerably less eccentric than those generated via the binear array of evaluation 1.

IV. OPTIMAL MICROPHONE PLACEMENT

As evaluations 1 and 2 illustrate, the placement of sensors within a room can dramatically affect the accuracy of the source-location estimates. Sensor positioning is generally subject to the physical or aesthetic constraints of the environment, but may also be restricted by the requirements of the time-delay estimation procedure. It has been assumed throughout this discussion that the source localization process is independent of the TDOA estimation, requiring only the sensor locations, TDOA estimates, and the estimate variances. However, the precision of time-delay estimators is
highly dependent upon the coherence or similarity of the signals received at the two sensors. It is therefore essential to the accuracy of the TDOA estimates that the separation of the sensors within each sensor pair be small enough to prevent significant disparities in the received signal quality or content between the sensor pair. This qualification makes certain placement scenarios that seem advantageous for localization, ineffective due to the detrimental effects on the accuracy of the time-delay estimates; the increased TDOA variance may overwhelm the advantages of a broadened baseline. In practice, the selection of sensor separation distances requires a knowledge of the environmentally dependent performance characteristics associated with the time-delay estimation procedure employed.

The choice of the number of sensor pairs and their positions ultimately depends upon minimizing some form of a precision-based cost function that is constrained by the number of available sensors, the physical environment, and the required intrapair separation distances. The details and method of minimizing such a cost function will vary from one application to another. Some work in this area related specifically to speech source acquisition has been reported in.3,7,38,39 In many scenarios, prior information concerning the potential locations of signal sources or a set of spatial regions from which it is desirable to obtain “good” location estimates may be specified and the complexity of the cost function will be greatly reduced. Regardless of the specifics, at the core of an optimization procedure there must be a means of evaluating estimation accuracy given source and sensor locations. The previously derived expressions for predicting error covariance may be applied for this purpose.

The following simulations exhibit the application of the error estimation procedure to array design in a videoconferencing environment. For this example the source positions and orientations are explicit and a practical cost function is easily defined. Furthermore, the simulation procedure detailed below may easily be modified and extended to cover a wide range of applications.
A. Video-conferencing simulation parameters

Details of the simulations are given below:

1. The room

A 4-m×1.2-m×0.1-m rectangular table at a height of 0.7 m, centered within a 4-m×7-m×2.75-m idealized, anechoic enclosure. See Fig. 5. No modeling of reverberations or secondary sources is included.

2. Talker locations

Participants are seated (height=1.25 m) and standing (height=1.75 m) at 0.75-m intervals along the sides of the table, at the table head, and a remote positions off of the corner locations for a total of 30 speaker locations. Associated with each speaker location is source orientation vector. In this case, all the talkers face a monitor/camera found at the front of the table and at a height of 1.5 m. Again, see Fig. 5.

3. Microphone placement constraints

Microphones are restricted to occur in four-element arrays each consisting of two orthogonal microphone pairs with 25-cm sensor separations as shown in Fig. 6. These particular sensor restrictions are dictated by the practical time-delay estimation considerations discussed above and by the requirements of the linear intersection location estimator, a closed-form localization algorithm detailed in Ref. 23 and employed for the actual location estimation in practice. The four-element arrays are further constrained to placements on the monitor/camera unit, the table surface, the room walls, or the table ceiling (microphones are assumed to be held in chandelier units and placed at a height 0.25 m below the actual ceiling). In each case the individual microphones are assumed to be mounted normal to the particular surface and a microphone orientation vector is included along with sensor location.

4. Source model

The sources are modeled as cardioid radiators. The source amplitude is a function of radiation angle and distance. In particular:

\[ A(s, m) \propto \left( \frac{1 + \cos(\theta_{sm})}{2r_{sm}} \right), \]

where \( \theta_{sm} \) is the angle generated by the source orientation vector and the source-microphone displacement vector, \( s-m \), and \( r_{sm} \) is the source-microphone separation, \( |s-m| \).

5. Microphone model

The microphones are modeled as possessing a cosine reception pattern. The reception amplitude is a function of source angle relative to the microphone orientation vector,

\[ B(m, s) \propto |\cos(\theta_{ms})|, \]

where \( \theta_{ms} \) is the angle generated by the microphone orientation vector and the source-microphone displacement vector.

6. TDOA variance

The variance of the TDOA estimate, \( \sigma^2 \), for a source \( s \) relative to the \( i \)th sensor pair \( \{m_1, m_2\} \) is assumed to be related to the sum of the reciprocal signal-noise power ratio at the two microphones:

\[ \sigma^2 = \text{variance}(\tau_i) \]

\[ \propto \left( \frac{1}{A(s, m_1)B(m_1, s)} \right)^2 + \left( \frac{1}{A(s, m_2)B(m_2, s)} \right)^2. \]

The noise power is assumed to be uniform at each microphone and is therefore incorporated into the constant of proportionality.

Given an arrangement of four-element microphone-arrays, the localization error covariance for a particular source-location may be calculated via Eq. (16). The localization error associated with a combination of microphone arrays, \( \mathbf{M} \), and the set of predefined source locations is the average of the total standard deviations for the 30 individual error covariances: Average total error \( \{\mathbf{M}\} \)

\[ = \frac{1}{30} \sum_{i=1}^{30} \text{trace}[\text{cov} \{\Delta s_{\mathbf{M}_i}\}]^{1/2}, \]

where \( \text{cov} \{\Delta s_{\mathbf{M}_i}\} \) is the error covariance associated with the \( i \)th source location. The units of this statistic are meters however, given the constants of proportionality in the modeling expressions, the total average error values presented will be normalized relative to the minimum error microphone positioning scenario [see Fig. 8(a)].

The first set of simulations considered various arrangements of two four-element arrays. In each case, the arrays were allowed to vary with two degrees of freedom along a
specified surface (the monitor, ceiling, table, or walls). Array positions were selected to minimize the average total standard deviation value associated with a particular set of surfaces. Figure 7 presents the best results obtained for several combinations of arrays. Each of the plots displays an overhead view of the room with the principal component vectors associated with each talker location included. For each scenario the principal components and the average total error listed have been scaled relative to a uniform noise condition.

Of the five scenarios considered, the monitor/ceiling arrangement, plot (A), yields the lowest overall error, 2.47. The best positioning of the two arrays was found when the monitor array was mounted on the center of the monitor unit at a height of 1.22 m, just below the focus point of the talker orientations. In the plot, the monitor array is illustrated as three small circles protruding from the edge of the monitor unit. The corresponding placement of the ceiling array is 0.25 m behind the table center as shown in plot (A). The principal components of the error covariance display behavior similar to the results of the simulations conducted in the previous section; the major components are generally oriented toward the array units and increase with their range from the arrays. Placing one array on the monitor and one on the table, plot (B), produced the next best results. For this case the height of the table array was compensated for by a positioning of the monitor mounted array at a height of 1.77 m, above the focus point. Comparing scenarios involving a monitor array, plots (A) and (B), to those without, plots (C), (D), and (E), it is apparent that it is preferable to have microphones at the room focal point rather than on the ceiling or table, even though arrays on the ceiling and table are closer to the candidate source locations. Finally, plot (D) incorporating two table arrays, yielded the worst performance of the conditions tested. Interestingly, the result for this condition is not symmetric. The rightmost table array consistently finds a minimal position 0.2 m from the table midline.

Figure 8 displays the results obtained with various combinations of three array units. In general, the inclusion of the third array greatly improves the overall localization accuracy relative to the two array scenarios. Once again, conditions involving a monitor array, plots (A), (B), (E), and (F), achieve the smallest error figures. The monitor/ two ceiling combination, plot (A), produces the best results of any condition evaluated, slightly outperforming the monitor/two table, plot (B), and monitor/table/ceiling, plot (F), scenarios. The three ceiling, (C), and three table, (D), situations are the least effective of the three array unit combinations and exhibit a curious asymmetry in their optimal placement; the rightmost array is approximately 0.2 m off of the midline in each case.

V. DISCUSSION

In this paper we have described a means for evaluating the error region associated with talker-location data obtained from a microphone-array system. The error predictor was shown to accurately model simulated source position estimates generated for a series of ideal room environments. Further experiments involving real rooms, physical talkers and actual microphone-array systems have been conducted and are given in Ref. 21. The results of these real-world evaluations confirm the results of the simulations presented here; the error covariance predictor serves as a valid indicator of the array source-localization performance.

In addition to providing a measure of confidence for the source location estimates, the error region predictor may be incorporated into a host of practical applications. As the preceding section illustrates, the predictor is a useful basis for a sensor placement error criteria and provides an effective tool for array design with regard to source location estimation in room environments.
