

AUTOMATIC POSITION CALIBRATION OF MULTIPLE MICROPHONES

Vikas C. Raykar and Ramani Duraiswami

Perceptual Interfaces and Realities Lab., University of Maryland, CollegePark

ABSTRACT

In this paper we describe a method to automatically determine the relative three dimensional positions of multiple microphones using at least five loudspeakers in unknown positions. The only assumption we make is that there is a microphone which is very close to a loudspeaker. In our experimental setup we attach one microphone to each loudspeaker. We derive the Maximum Likelihood estimator and the solution turns out to be a non-linear least squares problem. A closed form solution which can be used as a initial guess for the minimization routine is derived. We also derive an approximate expression for covariance of the estimator using implicit function theorem. Using this we analyze the performance of the estimator with respect to the positions of the loudspeakers. The algorithm is validated using both Monte-Carlo simulations and a real-time experimental setup.

1. INTRODUCTION

Microphone arrays are widely used for applications like sound source localization and tracking, hands free voice communication and speech enhancement. Most multi-microphone array processing algorithms need to know the positions of the microphones very precisely. In the case of source localization, even relatively small uncertainties in sensor location could make substantial, often dominant, contributions to overall localization error [1]. Most of the current system implementations place the microphones in known positions. However in ad-hoc deployed arrays it is rather tedious and very often not accurate to get the microphone positions manually using a tape or laser devices. Also the geometry of the array may change over time frequently either accidentally or due to redeployment. So automatic position calibration of multiple microphones is very essential. In this paper we propose a method to automatically determine the three dimensional positions of multiple microphones.

Following are the novel contributions of this paper: Our formulation for position calibration of microphones assumes that the positions of loudspeakers are unknown. As a result we do not need a precisely arranged setup of loudspeakers as in [2]. Previous work on position calibration with unknown source locations [1, 3] derive the solution as a non-linear minimization problem. However the numerical optimization methods do not converge unless we have a very close initial guess. We propose a closed form solution for the microphone and loudspeaker coordinates. We derive the approximate mean and covariance of the implicitly defined estimator using the implicit function theorem and Taylor series expansion. We analyze the calibration accuracy as a function of the position of the loudspeakers. In particular we show that the loudspeakers should be placed as far away from each other and all the microphones should be in the convex hull formed by the loudspeakers as opposed to the setup in [2] where all the loudspeakers are close to each other.

2. MAXIMUM LIKELIHOOD ESTIMATOR

Given a set of M microphones and S loudspeakers in unknown locations, our goal is to estimate their three dimensional coordinates. Each of the loudspeaker is excited using a known calibration signal such as maximum length sequence or chirp signal and the signal is captured by each of the microphones. The Time of Flight (TOF) is estimated from the captured audio signal. The TOF for a given pair of microphone and speaker is defined as the time taken by the acoustic signal to travel from the speaker to the microphone. Let $\mathbf{m}_i = [mx_i \ my_i \ mz_i]^T$ and $\mathbf{s}_j = [sx_j \ sy_j \ sz_j]^T$ be the three dimensional vectors representing the x, y and z coordinates of the i^{th} microphone and j^{th} loudspeaker respectively. Let $TOF_{ij}^{estimated}$ and TOF_{ij}^{actual} be the estimated and the actual TOF respectively for the i^{th} microphone and j^{th} speaker. The actual TOF can be written as

$$TOF_{ij}^{actual} = \frac{\|\mathbf{m}_i - \mathbf{s}_j\|}{c} \quad (1)$$

where $\|\cdot\|$ is the Euclidean norm and c is the speed of the sound in the acoustical medium¹. Assuming a Gaussian noise model for our observations we can derive the Maximum Likelihood (ML) estimator as follows. Let Θ , be a vector of length $P \times 1$, representing all the unknown non-random parameters to be estimated (microphone and speaker coordinates). Let Γ , be a vector of length $N \times 1$, representing noisy estimated TOF measurements. Let $T(\Theta)$, be a vector of length $N \times 1$, representing the actual value of the TOF observations. Then our model for the observations is $\Gamma = T(\Theta) + \eta$ where η is the zero-mean additive white Gaussian noise vector of length $N \times 1$ where each element has the variance σ_η^2 . Also let us define Σ to be the $N \times N$ covariance matrix of the noise vector η . The likelihood function of Γ in vector form can be written as

$$p(\Gamma/\Theta) = (2\pi)^{-\frac{N}{2}} |\Sigma|^{-\frac{1}{2}} \exp -\frac{1}{2}(\Gamma - T)^T \Sigma^{-1} (\Gamma - T) \quad (2)$$

The ML estimate $\hat{\Theta}_{ML}$ is that Θ which maximizes the likelihood ratio and can be shown to be

$$\hat{\Theta}_{ML}(\Gamma) = \arg\{\max_{\Theta} F(\Theta, \Gamma)\}$$

$$F(\Theta, \Gamma) = -\frac{1}{2}[\Gamma - T(\Theta)]^T \Sigma^{-1} [\Gamma - T(\Theta)] \quad (3)$$

Assuming that each of the TOFs are independently corrupted by zero-mean additive white Gaussian noise² of variance σ_{ij}^2 the

¹The speed of the sound depends on the room temperature and is given by $c = (331 + 0.6T)m/s$, where T is the temperature in degree Celsius. In practice we assume that c is known and constant. However we can also estimate the speed of the sound along with the positions of the microphones as in [2].

²We estimate the TOF using Generalized Cross Correlation (GCC)[4]. The estimated TOF is corrupted due to ambient noise and room reverberation. For high SNR the delays estimated by the GCC can be shown to be normally distributed with zero mean [4].

ML estimate can also be formulated as a nonlinear least squares problem (in this case Σ is a diagonal matrix), i.e.

$$\hat{\Theta}_{ML} = \arg_{\Theta} \min \sum_{i=1}^M \sum_{j=1}^S \frac{(TOF_{ij}^{estimated} - TOF_{ij}^{actual})^2}{\sigma_{ij}^2} \quad (4)$$

Since the solution depends only on pairwise TOFs, any translation, rotation and reflection of the global minimum found will also be a global minimum. In order to make the solution invariant to rotation and translation we select three arbitrary nodes to lie in a plane such that the first is at $(0, 0, 0)$, the second at $(x_1, 0, 0)$, and the third at $(x_2, y_2, 0)$. To eliminate the ambiguity due to reflection along Z-axis we specify one more node to lie in the positive Z-axis. Also the reflections along X-axis and Y-axis can be eliminated by assuming the nodes which we fix to lie on the positive side of the respective axes i.e $x_1 > 0$ and $y_2 > 0$. The ML estimate for the node coordinates of the microphones and loudspeakers is implicitly defined as the minimum of the non-linear function given in Equation 4. The Levenberg-Marquardt method is a popular method for solving non-linear least squares problems. One main problem is that the minimization routine will not converge to the global minima unless we have a very good initial guess. In section 3 we derive a closed form solution which can be used as a initial guess for the minimization routine. Also for least squares the total number of observations should be greater than or equal to the total number of parameters to be estimated. In our case $MS \geq 3(M + S) - 6$. If $M = S = K$ then $K \geq 5$.

3. CLOSED FORM SOLUTION

Given the pairwise Euclidean distances between N nodes their relative positions can be determined by means of metric or classical Multidimensional Scaling (MDS) [5]. Given a set of N points in three dimensional space, let X be a $N \times 3$ matrix where each row represents the 3D coordinates of each point. Then the $N \times N$ matrix $B = XX^T$ is called the dot product matrix. Starting with a matrix B (possibly corrupted by noise), it is possible to factor it to get the matrix of coordinates X . One method to factor B is to use singular value decomposition (SVD), i.e., $B = U\Sigma U^T$ where Σ is a $N \times N$ diagonal matrix of singular values. The diagonal elements are arranged as $s_1 \geq s_2 \geq s_r > s_{r+1} = \dots = s_N = 0$, where r is the rank of the matrix B . The columns of U are the corresponding singular vectors. We can write $X' = U\Sigma^{1/2}$. From X' we can take the first three columns to get X . If the elements of B are exact (i.e., they are not corrupted by noise), then all the other columns are zero.

In practice, we can estimate the distance matrix D , where the ij^{th} element is the Euclidean distance between the i^{th} and the j^{th} point. This distance matrix D must be converted into a dot product matrix B before MDS can be applied. We need to choose some point as the origin of our coordinate system in order to form the dot product matrix. Let us say we choose the k^{th} node as the origin of our coordinate system. Let d_{ij} and b_{ij} be the distance and dotproduct respectively, between the i^{th} and the j^{th} node. Referring to Figure 1, using the cosine law we have $d_{ij}^2 = d_{ki}^2 + d_{kj}^2 - 2d_{ki}d_{kj}\cos(\alpha)$. The dot product is given by $b_{ij} = d_{ki}d_{kj}\cos(\alpha)$. Combining the above two equations we get $b_{ij} = \frac{1}{2}(d_{ki}^2 + d_{kj}^2 - d_{ij}^2)$. Any point can be selected as the origin, but if the distances have random errors then choosing the centroid as the origin will minimize the errors as they tend to cancel each

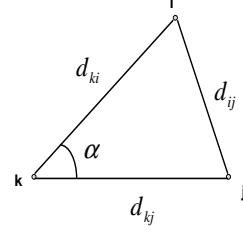


Fig. 1. Law of cosines

other. It can be shown that with respect to the centroid

$$b_{ij}^* = -\frac{1}{2} \left[d_{ij}^2 - \frac{1}{N} \sum_{l=1}^N d_{il}^2 - \frac{1}{N} \sum_{m=1}^N d_{mj}^2 + \frac{1}{N^2} \sum_{o=1}^N \sum_{p=1}^N d_{op}^2 \right]$$

In our case of M microphones and S speakers we cannot use MDS directly because we cannot measure all the pairwise distances. We only can measure the distance between each speaker and all the microphones. However we cannot measure the distance between two microphones or two speakers. In our practical setup for every loudspeaker we use we have a microphone attached to it. So we cluster microphones and speakers, which are close together as one entity i.e. we assume that the distance between them is zero. Based on this approximation, the distance d_{ij} between the i^{th} and j^{th} microphone-loudspeaker pair is given by

$$d_{ij} \approx 0 \text{ if } i = j \quad d_{ij} \approx \frac{c(TOF_{ij} + TOF_{ji})}{2} \text{ if } i \neq j \quad (5)$$

where c is the speed of the sound. Once we have all the pairwise distances we use classical MDS to get the approximate positions of the microphone-loudspeaker pairs. The position estimate from MDS is with respect to the centroid and the orientation arbitrary and hence it is converted into the reference coordinate system. The approximate locations of the microphone-loudspeaker pairs are slightly perturbed to get the initial guess for the microphone and speaker locations. We use this as an initial guess for the non-linear minimization routine and get the exact locations of the microphones and loudspeakers in each microphone-loudspeaker pair. As discussed before for the ML estimation procedure we need a minimum of five microphone-loudspeaker pairs.

From the previous step we know the location of five loudspeakers and the microphone close to them. If the location of four speakers are known then by trilateration the microphones position can be determined analytically. If the distance to more than four loudspeakers are known then we solve the problem in a least square sense. As before let us say we have S loudspeakers. Let $\mathbf{s}_j = [sx_j sy_j sz_j]^T$ be the x, y and z coordinates of the j^{th} speaker. Let $\mathbf{m}_i = [mx_i my_i mz_i]^T$ be the unknown microphone coordinates which we have to determine. For the i^{th} microphone we have S TOF measurements i.e. $c^2 TOF_{ij}^2 = \|\mathbf{m}_i - \mathbf{s}_j\|^2 \quad j = 1 \dots S$. In order to write a closed form solution for m_i we take the difference of every pairwise equations i.e.

$$\|\mathbf{m}_i - \mathbf{s}_j\|^2 - \|\mathbf{m}_i - \mathbf{s}_k\|^2 = c^2 TOF_{ij}^2 - c^2 TOF_{ik}^2 \quad (6)$$

This can be simplified to,

$$(\mathbf{s}_k - \mathbf{s}_j)^T \mathbf{m}_i = b_{jk}^i \quad (7)$$

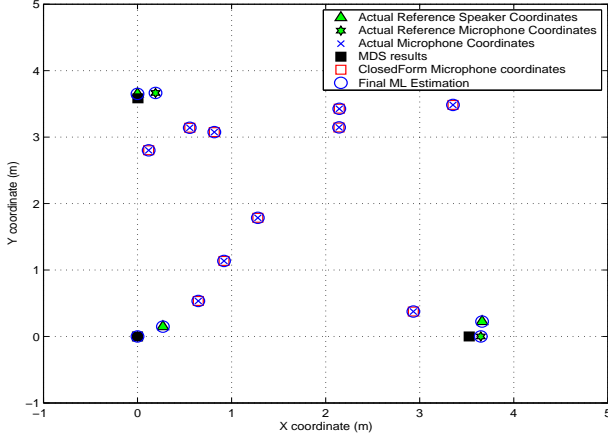


Fig. 2. Result of the proposed algorithm in two dimensions consisting of 10 microphones and 3 microphone-loudspeaker pairs.

where $b_{jk}^i = [c^2 TOF_{ij}^2 - c^2 TOF_{ik}^2 - \|s_j\|^2 + \|s_k\|^2]/2$. Each pair of speakers gives rise to one equation in 3 unknowns. If we have S speakers then we will have $S(S-1)/2$ equations. So we need 3 pairs i.e a minimum of 4 speakers to determine the position of one microphone. For S speakers we define the following matrix A and the vector \mathbf{b}^i

$$A = \begin{bmatrix} (s_1 - s_2)^T \\ \vdots \\ (s_{S-1} - s_S)^T \end{bmatrix} \mathbf{b} = \begin{bmatrix} b_{21}^i \\ \vdots \\ b_{S(S-1)}^i \end{bmatrix} A\mathbf{m}_i = \mathbf{b}^i \quad (8)$$

The least square solution can be written as $\mathbf{m}_i = (A^T A)^{-1} A^T \mathbf{b}^i$. The closed form solution for the microphone coordinates is further refined by doing a final ML estimation of all the parameters i.e. the microphones and the speaker positions.

Figure 2 shows an example in two dimensions with 10 microphones and 3 microphone-loudspeaker pairs. First using MDS we get the approximate locations of the three microphone-loudspeaker pairs which is shown as filled square in the figure. This approximate position is refined using ML estimation procedure to get the actual location of the microphone and the loudspeaker in the microphone-loudspeaker pair. Using the location of the loudspeakers we get a closed form solution for the microphone locations which are shown as squares. In the final ML estimation we refine the closed form solution to get the exact location of the microphones (shown as circles).

4. ESTIMATOR VARIANCE

The ML estimate for the microphone and speaker positions is defined implicitly as the minimum of a certain error function (refer Equation 4). Hence it is not possible to get exact analytical expressions for the mean and the variance of the estimator. However by using the implicit function theorem and the Taylor's series expansion it is possible to derive approximate expressions for the mean vector and covariance matrix of implicitly defined estimators [6, 7]. We give a brief outline of the derivation. The ML estimate $\hat{\Theta}$ of Θ is the one which maximizes the likelihood ratio. In our case the ML estimator is implicitly defined by Equation 3. The maximum can be found by setting the first derivative to zero i.e. $\nabla_{\Theta} F(\Theta, \Gamma) |_{\Theta=\hat{\Theta}} = \mathbf{0}$. The implicit function

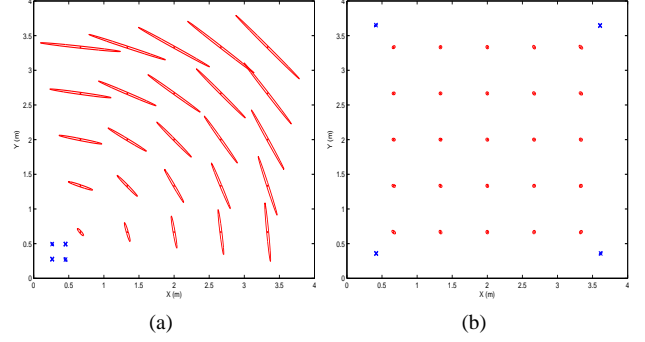


Fig. 3. 95% uncertainty ellipses for a regular 2 dimensional array of 25 microphones and 4 loudspeakers. Noise variance for all cases is $\sigma^2 = 10^{-9}$. The microphones are represented as dots (.) and the loudspeakers as (x). The position of one loudspeaker and the x coordinate of another is assumed to be known. In (a) loudspeakers are close to each other and in (b) they are spread out one at each corner of the grid.

theorem guarantees that this equation implicitly defines a vector valued function $\hat{\Theta} = h(\Gamma) = [h_1(\Gamma), h_1(\Gamma), \dots, h_P(\Gamma)]^T$ that maps the observation vector Γ to the parameter vector $\hat{\Theta}$. Hence $\nabla_{\Theta} F(h(\Gamma), \Gamma) = \mathbf{0}$. However it is not possible to find an analytical expression for $h(\Gamma)$. But we can approximate the covariance using the first-order Taylor series expansion for $h(\Gamma)$. Let Γ_m be the mean of Γ . Then expanding $h(\Gamma)$ around Γ_m we get

$$h(\Gamma) \approx h(\Gamma_m) + [\nabla_{\Gamma} h(\Gamma)^T |_{\Gamma=\Gamma_m}]^T (\Gamma - \Gamma_m) \quad (9)$$

Using this expression we get

$$Cov[h(\Gamma)] = [\nabla_{\Gamma} h(\Gamma)^T |_{\Gamma=\Gamma_m}]^T Cov(\Gamma) [\nabla_{\Gamma} h(\Gamma)^T |_{\Gamma=\Gamma_m}] \quad (10)$$

Note we do not know $h(\Gamma)$. Differentiating $\nabla_{\Theta} F(h(\Gamma), \Gamma) = \mathbf{0}$ with respect to Γ and evaluating it at Γ_m yields

$$\nabla_{\Theta} \nabla_{\Theta} F(h(\Gamma_m), \Gamma_m) [\nabla_{\Gamma} h(\Gamma_m)^T]^T + \nabla_{\Theta} \nabla_{\Gamma} F(h(\Gamma_m), \Gamma_m) = \mathbf{0} \quad (11)$$

Substituting for the derivatives we get $Cov[\hat{\Theta}] \approx [J^T \Sigma^{-1} J]^{-1}$ where J is a $N \times P$ matrix of partial derivatives of $T(\Theta)$ called the *Jacobian* of $T(\Theta)$ where each element is $[J]_{ij} = \frac{\partial T_i(\Theta)}{\partial \Theta_j}$. If we assume that all the microphone and source locations are unknown, F is rank deficient and hence not invertible. This is because the solution to the ML estimation problem as formulated is not invariant to rotation and translation. In order to make F invertible we remove the rows and columns corresponding to the known parameters. In our formulation we assumed that we know the positions of a certain number of nodes, i.e we fix three of the nodes to lie in the x - y plane. The covariance matrix depends on which of the sensor nodes are assumed to have known positions. Figure 3 shows the 95% uncertainty ellipses for a regular two dimensional array containing 25 microphones and 4 loudspeakers for different positions of the loudspeakers. In Figure 3(a) all the four loudspeakers are at one corner of the grid. The uncertainty in the direction tangential to the line joining the microphone and the center of the known nodes is much larger than along the line. It is beneficial if the known nodes are on the edges of the network and as far away from each other as possible. In Figure 3(b) the known loudspeakers are on the edges of the network. As can be seen there is a substantial reduction in the dimensions of the uncertainty ellipses.

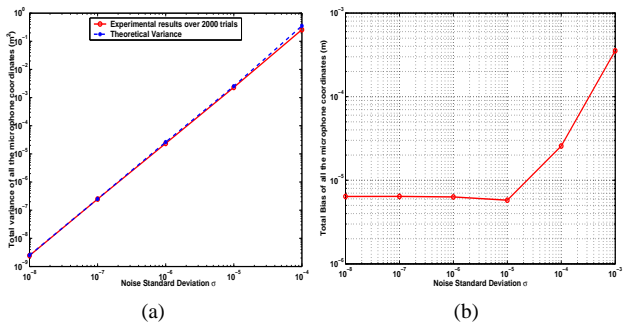


Fig. 4. (a) The total variance and (b) total bias of all the microphone coordinates for increasing noise standard deviation σ . The network consisted of 20 microphones and 5 microphone-loudspeaker speakers. The theoretical variance is also shown.

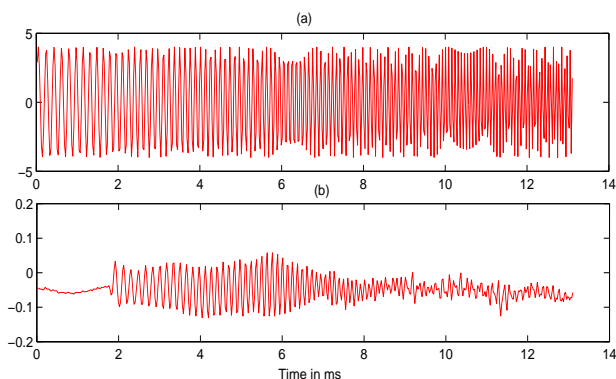


Fig. 5. (a) The actual chirp signal used in our setup and (b) the chirp signal received by a microphone.

In order to validate the experimental performance with the theoretical variance derived, we performed a series of Monte Carlo simulations with 20 microphones which were randomly selected to lie in a room of dimensions $4.0m \times 4.0m \times 4.0m$. Five microphone-loudspeaker pairs we placed such that all the 20 microphones were within the convex hull formed by the five pairs. Figure 4(a) and Figure 4(b) show the total variance (sum of all the estimated variances of each parameter) and the total bias (sum of all the estimate biases for each parameter) of all the unknown microphone coordinates plotted against the noise standard deviation σ . The theoretical variance is also shown. From Figure 4(b) we can see that the simulated results closely track the theoretical variance.

5. EXPERIMENTAL SETUP AND RESULTS

In order to measure the TOF we used a cosine chirp signal of 512 samples at 39.0625 kHz as our calibration signal. The instantaneous frequency varied linearly from 5 kHz to 10 kHz. Figure 5(a) shows the chirp signal as sent out by a speaker. Figure 5(b) shows the chirp signal recorded by a microphone. The distortion and the spreadout is due to the speaker, microphone and room response. In order to measure the time delay we used the Generalized Cross Correlation (GCC) method [4] with Phase Transform (PHAT) weighting. We have set up a 32 element microphone array as shown in Figure 6(a). In order to calibrate this array we placed the tiny speaker at five different positions on the array such that the speaker was very close to a microphone. The first four

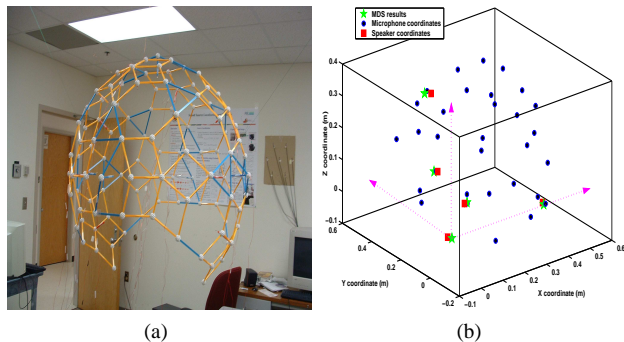


Fig. 6. (a) The 32 element microphone array and (b) the results obtained from our algorithm.

speakers define our coordinate system and all the positions are with respect to this coordinate system. Figure 6(b) shows the results obtained by our algorithm for this microphone array. The '*' indicates the positions of the speakers and the circles indicate the microphone positions.

6. CONCLUSIONS

An algorithm for automatic position calibration of multiple microphones is presented. Our method does not require the position of the loudspeakers to be known. The only constraint we impose is that we attach a microphone to a loudspeaker. We derived a closed form solution which was further refined by nonlinear minimization. We also derived the variance of our estimator and extensively validated the algorithms on simulated and real data.

7. REFERENCES

- [1] Y. Rockah and P. M. Schultheiss, "Array shape calibration using sources in unknown locations Part II: Near-field sources and estimator implementation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, no. 6, pp. 724–735, June 1987.
- [2] J. M. Sachar, H. F. Silverman, and W. R. Patterson III, "Position calibration of large-aperture microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002, pp. II-1797 – II-1800.
- [3] A. J. Weiss and B. Friedlander, "Array shape calibration using sources in unknown locations—a maximum-likelihood approach," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 12, pp. 1958–1966, December 1989.
- [4] 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, August 1976.
- [5] W. S. Torgerson, "Multidimensional scaling: I. theory and method," *Psychometrika*, vol. 17, pp. 401–419, 1952.
- [6] J. A. Fessler, "Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): Applications to tomography," *IEEE Trans. on Image Processing*, vol. 5, no. 3, pp. 493–506, March 1996.
- [7] A. R. Chowdhury and R. Chellappa, "Statistical bias and the accuracy of 3d reconstruction from video," *Submitted to International Journal of Computer Vision*.