3D BINAURAL AUDIO CAPTURE AND REPRODUCTION USING A MINIATURE MICROPHONE ARRAY

Shengkui Zhao[*]
Advanced Digital Science Center
Illinois at Singapore
Singapore
shengkui.zhao@adsc.com.sg

Ryan Rogowski, Reece Johnson and Douglas L. Jones
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Illinois, USA
jones@ifp.illinois.edu

ABSTRACT

This paper presents a new low-cost and efficient approach for the real-time three-dimensional (3D) binaural audio capture and reproduction via headphones using a miniature microphone array. The microphone array is configured in B-format to minimize space requirement using an omnidirectional microphone and three bidirectional microphones. The signals captured by the microphone array are applied by a set of optimal time-invariant gain vectors, which converts a B-format Ambisonic sinal into binaural signal for headphone reproduction. The optimal time-invariant gain vectors that are computed offline integrate the two stages of beamforming and head related transfer function (HRTF) filtering. As an alternative to the virtual speaker method, the proposed beamforming approach is independent of the number of virtual audio sources and flexible for working on different sets of HRTFs. A real-time system has been implemented based on the proposed method. Psychophysical hearing tests show good localization accuracy.

1. INTRODUCTION

Capture and reproduction of three-dimensional (3D) audio is becoming increasingly important for communication, virtual reality, and entertainment systems. The ultimate goal of a 3D binaural audio system is to accurately capture 3D sound and reproduce the 3D sound with spatial perception using speaker system. Although studies relating to reproduction technologies have been active, much work has concentrated on 3D binaural audio reproduction through loudspeakers [1]. Studies of 3D audio capture and reproduction through stereo headphones are far fewer.

In the literature, the *TeleHead* introduced in [2] employs a dummy head with placing two microphones in the left and right ear canals. Algazi et al. proposed a motion-tracked binaural recording technique called MTB [3]. In this approach, a sphere or a cylinder with several pairs of microphones was used. The above approaches, however, need to produce a personalized *TeleHead* to realize the sound space precisely for each listener in practice. Alternatively, several works have studied spherical microphone array (SMA) for spatial sound reproduction. M. Noisternig et al. [4] presented the virtual Ampbisonic approach for the playback of high order spatial audio (HOA) encoded sound fields using headphones. R. Duraiswami et al [5] and J. Meyer et al. [6] investigated beamforming SMA consisting of many microphones for the plane-wave decomposition of incident acoustic wave fields based on spherical harmonic transform (SHT). The incident plane-waves are then convolved with the head-related transfer function (HRTF) for the corresponding direction. To acquire 3D sound-space information without HRTF cues, S. Sakamoto et al. [7] proposed the SENZI system using 252 microphones mounted on a human-head-sized solid sphere and synthesized a 3D sound-space with high precision. All the above approaches are working properly in conditions. However, their systems are expensive, computationally demanding, and not very portable in general.

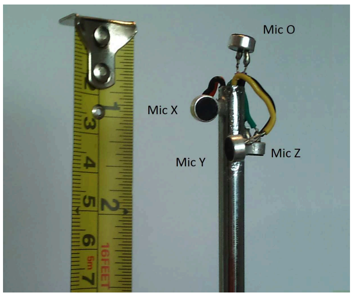


Figure 1: *Configure of the miniature XYZO array. The top microphone has an omnidirectional response and the bottom three microphones have gradient responses. Each sensor is* $6mm$ *(Diameter)* $\times 2.7mm$ *(Height)*

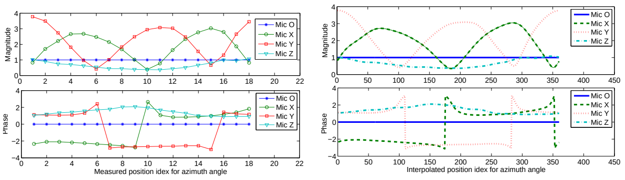
For a compact, low-cost system with good performance, in this paper we investigate the approach of creating a basis of functions using virtual cardioid steered to the different directions. Unlike the high-order SMA with large number of microphones [8], In this study, we use a low-order B-format microphone array [9]. By filtering each beam through the corresponding head-related transfer function (HRTF), human sound localization can be emulated based on the filtering effects of the human ear. Since the system requires only weighted combinations of the outputs of B-format microphones for forming fixed cardioid beams, the real-time processing is achieved.

2. THE B-FORMAT XYZO ARRAY AND 3D CALIBRATION

2.1. The B-Format XYZO array

The B-Format microphone array was first introduced by Michael Gerzon and Peter Fellgett in their research of Ambisonic [9] in the 1970th. Their goal was to have recordings of natural audio provid-

[*] This study is supported by the research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR).



(a) *Before interpolation* (b) *After interpolation*

Figure 2: *Frequency-domain response of the XYZO array at 1723 Hz for one round of azimuth angles at fix elevation angle.*

ing a spatial impression to the listener with 3-dimensional information in Multi Speaker surround sound. The Ambisonic technology of reproducing the 3-dimensional surround sound using loudspeaker signals is based on a spherical harmonic decomposition of the sound field corresponding to the sound pressure and the three components of the pressure gradient at a point space [10].

In this study, we use the microphone array configured in Fig. 1, which has three orthogonally mounted bidirectional pressure gradient microphones X, Y and Z, having figure-eight patterns with their directions of maximum response oriented in the X, Y, and Z axes, and one omnidirectional acoustic pressure microphone O for detecting sounds from all directions with equal magnitude. All the microphones are standard hearing-aid microphones that have useful directivity between 100 Hz and 20 kHz covering most sound sources. Each microphone measures only a few millimeters across, and is placed about a centimeter apart from the others. The array is named XYZO array in the paper. We virtually steer the XYZO array to the different directions of incident acoustic wave fields and then convolve the incident plane-waves with the HRTFs for the corresponding sound direction.

2.2. 3D Array Calibration

To reduce effects of phase mismatches and microphone positioning errors, a calibration process was performed. The impulse response of each microphone was measured through recording three continuous maximum-length sequences (MLS) signals playback by a loudspeaker [11]. The loudspeaker was placed approximately 2 m away from the XYZO array. The height of the loudspeaker was adjustable such that the measurements were conducted in the directions of azimuth angles with 20° resolution and elevation angles with 30° resolution around the XYZO array. By taking the circular cross-correlation between the microphone output and the MLS, the impulse responses were extracted. The frequency-domain steering vectors that are functions of direction and frequency were obtained by Fourier transforms.

The measured resolution of the steering vectors is usually insufficient to cover the full spatial directions in the three-dimensional space. Therefore, an interpolation approach is required to increase the measured resolution from 20° azimuth and 30° elevation to the resolutions of smaller angles. Using the two-dimensional Fourier-series, the measured steering vector for each microphone m were modeled as

$$e_m(\omega,\theta,\phi)=\sum_{p=-P}^{P}\sum_{q=-Q}^{Q}c_{p,q}(\omega)e^{-ip\theta}e^{-iq\phi} \qquad (1)$$

where the order of the Fourier series is $P\times Q$. By inserting the values of the measured steering vectors the coefficients of the Fourier series $c_{p,q}(\omega)$ were obtained as the least squares solution to an over-determined system. After solving for $c_{p,q}(\omega)$, the expansion was computed at higher resolution. This interpolation technique was successfully used for one-dimensional sound localization in [12]. Fig. 2 shows an interpolation result based on the two-dimensional Fourier-series fitting.

3. 3D AUDIO REPRODUCTION WITH XYZO ARRAY

In this section, we reproduce the 3D binaural audio by mapping the four-channel array outputs to the two-channel headphone. Based on the output model of the XYZO array, the beamforming theory, and the use of head related transfer function (HRTF), we derive the optimal gain vectors that can easily computed offline and applied to the array output in real-time.

3.1. Derivation of Optimal Gain Vectors

3.1.1. Modeling the Left-Ear and Right-Ear Signals

We now model the left-ear and right-ear signals for 3D binaural audio perception. By applying an optimal gain vector to the array outputs, we will show that the left-ear and right-ear signals can be well approximated.

Assuming that the HRTF data are static, the audio sources are far-field from the array, the audio signals at the left ear can be represented as

$$y_L(t)=\sum_{\boldsymbol{\theta}}h^L_{HRTF}(t,\boldsymbol{\theta})\otimes h_{RIR}(t,\boldsymbol{\theta})\otimes s(t,\boldsymbol{\theta}) \qquad (2)$$

where $\otimes$ denotes the convolution operation, $y_L(t)$ denotes the signals at the left ear. $h^L_{HRTF}(t,\boldsymbol{\theta})$ represents the time-domain left-ear HRTFs at the direction $\boldsymbol{\theta}\triangleq[\theta,\phi]$, where $\theta\in(0,360°]$ denotes the azimuth angle and $\phi\in[0,180°]$ the elevation angle in discrete values. $h_{RIR}(t,\boldsymbol{\theta})$ denotes the room impulse response for the audio source $s(t,\boldsymbol{\theta})$. Here we only show the derivations for the left-ear signal. The derivations for the right-ear signal can be obtained similarly.

In the Fourier transform domain, the left-ear signal can be represented as

$$Y(\omega)=\sum_{\boldsymbol{\theta}}H_{HRTF}(\omega,\boldsymbol{\theta})H_{RIR}(\omega,\boldsymbol{\theta})S(\omega,\boldsymbol{\theta}) \qquad (3)$$

where $Y(\omega)$, $H_{HRTF}(\omega,\boldsymbol{\theta})$, $H_{RIR}(\omega,\boldsymbol{\theta})$ and $S(\omega,\boldsymbol{\theta})$ are Fourier transforms of $y_L(t)$, $h^L_{HRTF}(\boldsymbol{\theta})$,$h_{RIR}(t,\boldsymbol{\theta})$and $s(t,\boldsymbol{\theta})$, respectively. The ω is the frequency index. We define $S_{RIR}(\omega,\boldsymbol{\theta})\triangleq H_{RIR}(\omega,\boldsymbol{\theta})S(\omega,\boldsymbol{\theta})$. The left-ear signal is then given in vector form as

$$Y(\omega)=\mathbf{h}^H_{HRTF}(\omega)\mathbf{s}_{RIR}(\omega) \qquad (4)$$

where H denotes the Hermitian transpose. The vector elements of $\mathbf{h}_{HRTF}(\omega)$ and $\mathbf{s}_{RIR}(\omega)$ are the $H_{HRTF}(\omega,\boldsymbol{\theta})$ and $S_{RIR}(\omega,\boldsymbol{\theta})$ ordered with the increasing values of θ and ϕ.

For the XYZO array, the signal received by each microphone can be modeled as

$$x_m(t)=\sum_{\boldsymbol{\theta}}e_m(\boldsymbol{\theta})\otimes h_{RIR}(t,\boldsymbol{\theta})\otimes s(t,\boldsymbol{\theta}) \qquad (5)$$

where $e_m(\boldsymbol{\theta})$ is the response of the mth microphone. Applying the Fourier transform on (5), we have

$$X_m(\omega)=\sum_{\boldsymbol{\theta}}E_m(\omega,\boldsymbol{\theta})S_{RIR}(\omega,\boldsymbol{\theta})=\mathbf{e}^H_m(\omega)\mathbf{s}_{RIR}(\omega) \qquad (6)$$

where $X_m(\omega)$ and $E_m(\omega, \boldsymbol{\theta})$ are Fourier transforms of $x_m(t)$ and $e_m(\boldsymbol{\theta})$, respectively, and the elements of $\mathbf{e}_m(\omega)$ and $\mathbf{s}_{RIR}(\omega)$ are $E_m(\omega, \boldsymbol{\theta})$ and $S_{RIR}(\omega, \boldsymbol{\theta})$ listed with the same order as $\mathbf{h}_{HRTF}(\omega)$ and $\mathbf{s}_{RIR}(\omega)$ in (4).

Comparing (4) and (6), at each frequency bin the left-ear signal $Y(\omega)$ can be approximated by applying a gain vector on the array output vector

$$\hat{Y}(\omega) = \mathbf{g}^H(\omega)\mathbf{x}(\omega) \tag{7}$$

where $\mathbf{g}(\omega) = [G_{L,X}(\omega), G_{L,Y}(\omega), G_{L,Z}(\omega), G_{L,O}(\omega)]^H$, and $\mathbf{x}(\omega) = [X_X(\omega), X_Y(\omega), X_Z(\omega), G_{L,O}(\omega)]^H$. We now need to derive the optimal gain vector $\mathbf{g}(\omega)$.

3.1.2. MVDR Beamforming Approach

The minimum-variance-distortionless-response (MVDR) frequency-domain beamformer maximizes the array output power at direction $\boldsymbol{\theta}$ using the following optimal weight vector [13], [14]:

$$\mathbf{w}_{opt}(\omega, \boldsymbol{\theta}) = \frac{\mathbf{R}_x^{-1}(\omega)\mathbf{e}(\omega, \boldsymbol{\theta})}{\mathbf{e}^H(\omega, \boldsymbol{\theta})\mathbf{R}_x^{-1}(\omega)\mathbf{e}(\omega, \boldsymbol{\theta})} \tag{8}$$

where $\mathbf{R}_x(\omega) \triangleq E\left[\mathbf{x}(\omega)\mathbf{x}^H(\omega)\right]$ is the correlation matrix of $\mathbf{x}(\omega)$ and $E[\cdot]$ stands for the expected-value operation. The steering vector is $\mathbf{e}(\omega, \boldsymbol{\theta}) \triangleq [e_X(\omega, \boldsymbol{\theta}), e_Y(\omega, \boldsymbol{\theta}), e_Z(\omega, \boldsymbol{\theta}), e_O(\omega, \boldsymbol{\theta})]^H$. Apply the HRTF $H_{HRTF}(\omega, \boldsymbol{\theta})$ to the corresponding beam of the MVDR beamformer and we have

$$\hat{Y}_{MVDR}(\omega) = \sum_{\boldsymbol{\theta}} H_{HRTF}(\omega, \boldsymbol{\theta}) \frac{\mathbf{e}^H(\omega, \boldsymbol{\theta})\mathbf{R}_x^{-1}(\omega)\mathbf{x}(\omega)}{\mathbf{e}^H(\omega, \boldsymbol{\theta})\mathbf{R}_x^{-1}(\omega)\mathbf{e}(\omega, \boldsymbol{\theta})} \tag{9}$$

According to (7), the optimal gain vector is then defined as

$$\mathbf{g}_{MVDR}(\omega) = \sum_{\boldsymbol{\theta}} H_{HRTF}(\omega, \boldsymbol{\theta}) \frac{\mathbf{e}^H(\omega, \boldsymbol{\theta})\mathbf{R}_x^{-1}(\omega)}{\mathbf{e}^H(\omega, \boldsymbol{\theta})\mathbf{R}_x^{-1}(\omega)\mathbf{e}(\omega, \boldsymbol{\theta})} \tag{10}$$

Considering the case of independent and identical distribution (IID) uncorrelated input sources with unit variance, we have

$$\mathbf{R}_x^{-1}(\omega) = \left(\mathbf{E}(\omega)\mathbf{R}_s(\omega)\mathbf{E}^H(\omega)\right)^{-1} \tag{11}$$

where $\mathbf{R}_s \triangleq E\left[\mathbf{s}_{RIR}(\omega)\mathbf{s}_{RIR}^H(\omega)\right]$ is the covariance matrix of the source signals with reverberation considered, and $\mathbf{E}(\omega) \triangleq [\mathbf{e}_W(\omega), \mathbf{e}_X(\omega), \mathbf{e}_Y(\omega), \mathbf{e}_Z(\omega)]^H$ is the steering matrix in the three-dimensional space. For simplicity, we assume that there is negligible reverberation in environment, and $\mathbf{R}_s(\omega)$ can be approximated as an identity matrix. The gain vector in (10) therefore can be simplified as

$$\mathbf{g}_{MVDR}(\omega) = \sum_{\boldsymbol{\theta}} \frac{\left(\mathbf{E}(\omega)\mathbf{E}^H(\omega)\right)^{-1}\mathbf{e}(\omega, \boldsymbol{\theta})H_{HRTF}(\omega, \boldsymbol{\theta})}{\mathbf{e}^H(\omega, \boldsymbol{\theta})\left(\mathbf{E}(\omega)\mathbf{E}^H(\omega)\right)^{-1}\mathbf{e}(\omega, \boldsymbol{\theta})} \tag{12}$$

Ideally, for the frequency independent microphones, equation (12) can be rewritten as

$$\mathbf{g}_{MVDR}(\omega) = \mu\left(\mathbf{E}(\omega)\mathbf{E}^H(\omega)\right)^{-1}\mathbf{E}(\omega)\mathbf{h}_{HRTF}(\omega) \tag{13}$$

where $\mu > 0$ is a scalar. The gain vector in (13) combines the beamforming with the steering matrix and the filtering of HRTFs. It is shown that by applying the left-ear and right-ear HRTFs for each frequency band, the gain vector for the left- and right-ear signals can easily be computed.

3.1.3. Minimum-Mean-Squared-Error (MMSE) Approach

Alternatively, we can solve for the gain vector $\mathbf{g}(\omega)$ from the following minimum-mean-squared-error (MMSE) estimation problem

$$\hat{\mathbf{g}}(\omega) = \underset{\mathbf{g}(\omega)}{argmin}\, E\left[\left|Y(\omega) - \hat{Y}(\omega)\right|^2\right] \tag{14}$$

Inserting (4), (6) and (7) into (14), $\hat{\mathbf{g}}(\omega)$ is given by

$$\underset{\mathbf{g}(\omega)}{argmin}\, E\left[\left|\mathbf{h}_{HRTF}^H(\omega)\mathbf{s}_{RIR}(\omega) - \mathbf{g}^H(\omega)\mathbf{E}(\omega)\mathbf{s}_{RIR}(\omega)\right|^2\right] \tag{15}$$

Solving (15), the optimal gain vector is obtained as

$$\hat{\mathbf{g}}_{opt}(\omega) = \left(\mathbf{E}(\omega)\mathbf{R}_s(\omega)\mathbf{E}^H(\omega)\right)^{-1}\mathbf{E}(\omega)\mathbf{R}_s(\omega)\mathbf{h}_{HRTF}(\omega) \tag{16}$$

After neglecting the reverberation, the covariance matrix is simplified to be an identity matrix, and the optimal gain vector is simplified as

$$\hat{\mathbf{g}}_{opt}(\omega) = \left(\mathbf{E}(\omega)\mathbf{E}^H(\omega)\right)^{-1}\mathbf{E}(\omega)\mathbf{h}_{HRTF}(\omega) \tag{17}$$

Comparing the gain vector in (13) and in (17), we can observe that the MVDR beamforming approach and the MMSE approach have equivalent forms. All the optimal gain vectors can be computed in an offline manner for the real-time system.

3.2. Real System Implementation

A real 3D audio capture and reproduction system has been implemented. The system has three units: signal acquisition unit, platform for 3D audio reproduction and 3D audio playback unit. The signal acquisition unit includes the components of the XYZO array, a signal pre-amp circuit and an M-Audio multiple-input multiple-output (MIMO) audio card. The The platform for 3D audio reproduction is a computer with Intel Xeon 2.67GHz CPU with 3GB memory. The 3D audio playback unit uses a stereo headphone connected to the output of the M-Audio audio card. The system was used to conduct the following experiments.

4. EXPERIMENTAL RESULTS

To evaluate the 3D localization accuracy of the proposed system, psychological subjective tests were carried out as follows: audio signals were 44.1 kHz, 16 bits acoustic signals that were abstracted in audio CD and recorded using the XYZO array. The HRTF database was measured using KEMAR dummy head at MIT Media Lab. The resolution of the HRTF measurements was interpolated the same way as we manipulated for the steering vector. There were five subjects involved in the evaluation. The assessment criterion is the ability to localize audio sources.

In the experiment, the tested audio sources were recorded as follows: for 0° elevation angle there are 24 positions recorded for the azimuth angles varying from 0° to 360° with 15° interval, and for 0° azimuth angle there were 24 positions recorded for the elevation angles varying from $-90°$ to 90°. The azimuth localization and elevation localization were tested separately. During each test, the subject was required to hear the processed audio signals through a stereo headphone and point to the direction of the audio source that was heard. A score of accuracy over 24 positions was obtained by comparing to the original directions of the audio

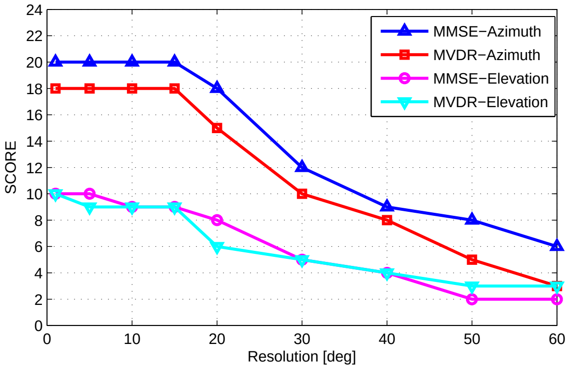


Figure 3: *Subjective sound localization test for variation of resolution.*

sources. The average scores over all the subjects were shown in Fig. 3. In addition, Fig. 3 also shows the evaluated results with different number of beams by changing the interpolated resolution as 1°, 5°, 10°, 15°, 20°, 30°, 40°, 50°, 60°. It was observed that when the resolutions were high, the average scores were also high. When the interpolated resolution drops to 60°, the scores was the lowest. Therefore, we concluded that using more beams will produce better localization. In addition, the sound localization for the azimuth positions was better than that of the elevation positions. By comparison, we convolved the playback signals with the HRTF database to obtain the approximated recordings of the KEMAR dummy head. The average scores over all the subjects were 21 for 24 azimuth positions and 10 for 24 elevation positions. Therefore, our proposed approach produces equivalent localization accuracy but with high portability.

Basically, the use of non-individual HRTF may create confusion in up/down and front/back directional perception. A compromised method of solving this problem is to use the personalized HRTFs or the general HRFTs averaged on the measurements with many subjects. In addition, head rotation may be taken into account with simple time-variant rotation matrices using a head tracker mounted on the headphones.

5. CONCLUSIONS

In this paper, we presented an approach for low-cost and efficient 3D binaural audio capture and reproduction using four coincident microphones. Besides the B-format microphone array, the method of obtaining the optimal gain vector can also be applied to other array configurations. The approach allows choosing the number of beams flexibly independent of the number of virtual sources. A subjective evaluation showed good localization accuracy.

6. REFERENCES

[1] M. Poletti, “A unified theory of horizontal holographic sound systems,” *Journal of Audio Engineering Society*, vol. 48, December 2000.

[2] I. Toshima, H. Uematsu, and T. Hirahara, “A steerable dummy head that tracks three-dimensional head movement: Telehead,” *Acoustical Science and Technogloy*, vol. 24(5), pp. 327–329, 2003.

[3] V.R. Algazi, R.O. Duda, and D.M. Thompson, “Motion-tracked binaural sound,” *Journal of Audio Engineering Society*, vol. 52(11), pp. 1142–1156, 2004.

[4] M. Noisternig, T. Musil, A. Sontacchi, and R. Holdrich, “3D binaural sound reproduction using a virtual Ambisonic approach,” *IEEE Int. Conf. on Virtual Environments, Human-Computer Interfaces, and Measurement Systems*, vol. 48, December 2000.

[5] R. Duraiswami, D. N. Zotkin, Z. Li, E. Grassi, N. A. Gumerov, and L. S. Davis, “High order spatial audio capture and its binaural head-tracked playback over headphones with HRTF cues,” *The 119th Audio Engineering Society Convention*, New York, USA, October 2000.

[6] J. Meyer, and T. Agnello, “Spherical microphone array for spatial sound recording,” *The 115th Audio Engineering Society Convention*, New York, USA, October 2003.

[7] S. Sakamoto, J. Kodama, and S. Hongo *et al*, “A 3d sound-space recording system using spherical microphone array with 252ch microphones,” *Proceedings of 20th International Congress on Acoustics*, August 2010.

[8] A. Farina, M. Binelli, A. Capra, and C. Varani, “Spatial analysis of room impulse reponses captured with a 32-capsules microphone array,” *The 130th Audio Engineering Society Convention*, London, UK, 2011.

[9] P.G. Craven and M.A. Gerzon, “Coincident microphone simulation covering three dimensional space and yielding various directional outputs,” *United States Patent*, vol. US 4,042, pp. 779, 1977.

[10] M. A. Gerzon, “The design of precisely coincident microphone arrays for stereo and surround sound,” *Audio Engineering Society Preprints*, February 1975.

[11] M. Lockwood and D.L. Jones, “Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms,” *Journal of Acoustic Society of America*, vol. 115, pp. 379–391, 2004.

[12] S. Mohan, M. Lockwood, M.L Kramer, and D.L. Jones, “Localization of multiple acoustic sources with small arrays using a coherence test,” *Journal of Acoustic Society of America*, vol. 123, pp. 2136–2147, 2006.

[13] J. Capon, “High-resolution frequency-wavenumber spectrum analysis,” *Proc. IEEE*, vol. 57(8), pp. 1408–1419, 1969.

[14] M. Lockwood and D.L. Jones *et al*, “Beamforming with collocated microphone arrays,” *Journal of Acoustic Society of America*, vol. 114, pp. 2451, 2003.