# OPTIMAL 3D BEAMFORMING USING MEASURED MICROPHONE DIRECTIVITY PATTERNS

*Mark R. P. Thomas, Jens Ahrens and Ivan Tashev*

Microsoft Research, One Microsoft Way, Redmond WA 98052, USA
{markth, jeahrens, ivantash}@microsoft.com

## ABSTRACT

The design of time-invariant beamformers is often posed as an optimization problem using practical design constraints. In many scenarios it is sufficient to assume that the microphones have an omnidirectional directivity pattern, a flat frequency response in the range of interest, and a 2D environment in which wavefronts propagate as a function of azimuth angle only. In this paper we consider a generalized solution for those cases in which one or more of these assumptions do not hold, yielding a beamformer that is optimized on measured directivity patterns as a function of azimuth, elevation and frequency. A comparative study is made with the 4-element cardioid microphone array employed in Microsoft Kinect for Windows, whose beamformer weights are calculated with directivity patterns using (a) 2D cardioid models, (b) 3D cardioid models and (c) 3D measurements. Results on a recorded noisy speech corpus show similar PESQ and speech recognition accuracy comparing (a) and (b), but a 50% relative improvement in word error rate using measured directivity patterns.

***Index Terms***— Microphone array, beamformer, superdirective beamformer, MVDR

## 1. INTRODUCTION

A microphone array is a device that samples a soundfield at multiple spatial locations, and whose output can be combined using a linear filterbank called a beamformer [1]. Beamformers are a class of spatial filters that are designed to improve the extraction of a wanted source signal compared with a single microphone, subject to certain design constraints. Adaptive data-dependent beamformers continually optimize their design based upon the signal and noise conditions [1]; in contrast, time-invariant beamformers make prior assumptions about their operating environment. Of the time-invariant approaches, superdirective beamforming [1] is desirable due to its ability to achieve high directivity with small apertures [2]. The Minimum Variance Distortionless Response (MVDR) beamformer can be designed for both time-invariant and adaptive cases by estimating the expected noise cross-power density to minimize the noise power at the beamformer output [3, 4]. Such designs can be sensitive to uncorrelated sensor self-noise and sensor mismatch, for which explicit constraints on the beamformer's white noise gain have been shown to successfully reduce such effects [1].

Beamformer designs often assume omnidirectional microphone directivity patterns with a flat frequency response. In real-world scenarios, physical factors due to microphone design constraints and the mounting hardware can have a significant effect upon the resulting directivity pattern and render it a function of azimuth, elevation and frequency. Beamformer design for arbitrary 2D directivity patterns and frequency responses has been considered in [5], accounting also for ambient and instrumental noise spectra to yield more realistic design. In this paper we formulate a generalized 3D solution and investigate the performance of an optimal beamformer by comparing designs based upon measured 3D directivity patterns and standard microphone models. Such a beamformer requires no additional computational overhead as the design modifies the steering weights only.

The remainder of this paper is organized as follows. In Section 2 the beamformer problem is formulated and an optimal solution is proposed based upon the MVDR criterion. In Section 3, beamformers are designed for a 4-channel microphone array and speech recognition error rates are discussed for each case. Concluding remarks are given in Section 4.

## 2. OPTIMAL BEAMFORMING

### 2.1. Problem Formulation

Let there be an array of microphones positions $p_m,\ m = 1, 2, \ldots, M$, where $p_m$ is a cartesian triplet $(x_m,\ y_m,\ z_m)$ in meters. For simplicity, it is assumed that all microphones are orientated with main response axis $\Omega_i = (0, 0)$ such that the 3D directivity pattern for an impinging wave from direction $\Omega = (\theta,\ \phi)$ is $U_m(f, \Omega)$, where $\theta = [-\pi/2, \pi/2]$ and $\phi = [0, 2\pi)$ are elevation and azimuth angles respectively. The midpoint of the array is placed at the origin of the coordinate system. Let $S_0(f)$ be a farfield source in the frequency domain located at angle $\Omega_0$. The response of the array is

$$\mathbf{X}(f) = \mathbf{D}_0(f)S_0(f) + \mathbf{N}(f), \tag{1}$$

where $\mathbf{X}(f) = [X_1(f)\ X_2(f) \dots X_M(f)]^T$ is an observation vector, $\mathbf{N}(f) = [N_1(f)\ N_2(f) \dots N_M(f)]^T$ is a noise vector and $\mathbf{D}_0(f) = [D_1(f)\ D_2(f) \dots D_M(f)]^T$ is a capture vector whose elements are

$$D_m(f) = e^{-j2\pi f \tau_m(\Omega_0)} U_m(f, \Omega_0). \tag{2}$$

The term $\tau_m$ accounts for the delay of the incoming wavefront at the $m$th sensor relative to the centre of the array. Similarly, the capture vector $\mathbf{G}(f, \Omega) = [G_1(f, \Omega)\ G_2(f, \Omega) \dots G_M(f, \Omega)]^T$ is defined for a general incidence angle $\Omega$,

$$G_m(f, \Omega) = e^{-j2\pi f \tau_m(\Omega)} U_m(f, \Omega). \tag{3}$$

Given observations $\mathbf{X}(f)$, the output of a generalized filter-and-sum beamformer is a weighted sum of the observations at each frequency bin [6],

$$Y(f) = \mathbf{W}_0^T(f)\mathbf{X}(f), \tag{4}$$

where $\mathbf{W}_0^T(f)$ is an $M \times 1$ vector of complex weights computed to steer the beam in the look direction $\Omega_0$. The resulting directivity pattern is a weighted sum of the capture vector elements at angle $\Omega$,

$$B(f, \Omega) = \mathbf{W}_0^T(f)\mathbf{G}(f, \Omega), \tag{5}$$

which, in the special case $\Omega = \Omega_0$,

$$B(f, \Omega_0) = \mathbf{W}_0^T(f)\mathbf{D}_0(f). \tag{6}$$

The aim is to design weights $\mathbf{W}_0^T(f)$ to form a beamformer subject to certain optimization criteria.

### 2.2. Calculation of Steering Weights

The design approach employed in this paper is based on the minimum variance distortionless response (MVDR) beamformer in the frequency domain [4]. We assume free-field propagation and that all sources lie in the farfield. Under ideal no-noise conditions, the beamformer output should equal the source signal such that $Y(f) = S(f)$. Additionally we aim to minimize the estimated noise variance. Combining (1) and (4) gives a new expression for the beamformer output,

$$Y(f) = \mathbf{W}_0^T(f)\mathbf{D}_0(f)S_0(f) + \mathbf{W}_0^T(f)\mathbf{N}(f) = S(f) + Y_N(f), \tag{7}$$

where $Y_N(f)$ is a noise term whose expected energy is [4]

$$Q = E[|Y_N(f)|^2] = \mathbf{W}_0^H(f)\mathbf{\Phi}_{NN}(f)\mathbf{W}_0(f), \tag{8}$$

where $(\cdot)^H$ denotes the conjugate transpose and $\mathbf{\Phi}$ is the noise cross-power spectral matrix:

$$\mathbf{\Phi}_{NN}(f) = \mathbf{N}(f)\mathbf{N}^H(f) = \begin{pmatrix} \Phi_{11} & \Phi_{12} & \dots & \Phi_{1M} \\ \Phi_{21} & \Phi_{22} & \dots & \Phi_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{M1} & \Phi_{M2} & \dots & \Phi_{MM} \end{pmatrix}. \tag{9}$$

Given known capture vectors $G_i(f, \Omega)$ and $G_j(f, \Omega)$, the elements of this matrix can be estimated assuming a spatially homogeneous and isotropic noise field by [7]

$$\Phi_{ij}(f) = N_0(f) \frac{N_{ij}(f)}{\sqrt{\bar{G}_i(f)\bar{G}_j(f)}}, \tag{10}$$

where $N_0(f)$ is the ambient noise spectrum and

$$N_{ij}(f) = \int_\Omega G_i(f, \Omega) G_j^*(f, \Omega) \mathrm{d}\Omega \tag{11}$$

$$\bar{G}_i(f) = \int_\Omega |G_i(f, \Omega)|^2 \mathrm{d}\Omega \tag{12}$$

$$\bar{G}_j(f) = \int_\Omega |G_j(f, \Omega)|^2 \mathrm{d}\Omega. \tag{13}$$

In a 2D scenario, the integrals are evaluated over azimuth angles in the interval $[0, 2\pi]$; in 3D, they are evaluated over all angles in $S^2$. This constrained minimization problem can be solved providing $N_0(f)$, $G_m(f, \Omega)$ and $p_m$ are known, either by imposing models or using measured data. Such a design is however sensitive to instrumental noise, particularly in the lower part of the frequency band. Without appropriate modification of the design criteria, the suppression of the ambient noise can be replaced by the amplified microphone self-noise leading to a non-robust solution. An additional term is therefore added to the $\mathbf{\Phi}_{NN}(f)$ to improve robustness [2]:

$$\mathbf{\Phi}_{N'N'}(f) = \mathbf{\Phi}_{NN}(f) + \mathbf{\Phi}_{\mathrm{II}}(f), \tag{14}$$

where $\mathbf{\Phi}_{\mathrm{II}}(f) = \kappa |N_{\mathrm{I}}(f)|^2 \mathbf{I}$ regularizes $\mathbf{\Phi}_{N'N'}(f)$ by accounting for uncorrelated instrumental noise with spectrum $N_{\mathrm{I}}(f)$, $\kappa$ is a regularization parameter and $\mathbf{I}$ is an $M \times M$ identity matrix. In practice this lowers the directivity index but increases the total noise suppression. The design procedure is summarized as a constrained optimization problem:

$$\begin{aligned} \widehat{\mathbf{W}}_0(f) &= \underset{\mathbf{W}_0(f)}{\arg\min} \mathbf{W}_0^H(f)\mathbf{\Phi}_{N'N'}(f)\mathbf{W}_0(f) \\ &\text{subject to } \mathbf{W}_0^T(f)\mathbf{D}_0(f) = 1. \end{aligned} \tag{15}$$

The linear constraint $\mathbf{W}_0^T(f)\mathbf{D}_0(f) = 1$ ensures a distortionless response in the steering direction. A closed form solution is given in the form [4]

$$\widehat{\mathbf{W}}_0(f) = \frac{\mathbf{D}_0^H(f)\mathbf{\Phi}_{N'N'}^{-1}(f)}{\mathbf{D}_0^H(f)\mathbf{\Phi}_{N'N'}^{-1}(f)\mathbf{D}_0(f)}, \tag{16}$$

which, in the extreme case where $\mathbf{\Phi}_{\mathrm{II}}(f) \gg \mathbf{\Phi}_{NN}(f)$, equates to the weights of a delay-and-sum beamformer

$$\widehat{\mathbf{W}}_0(f) = \frac{1}{M\mathbf{D}_0(f)}. \tag{17}$$

The practical implementation assumes an isotropic noise field making $\mathbf{\Phi}_{NN}(f)$ straightforward to estimate. These weights may be used to initialize an adaptive beamformer that continually updates the noise correlation matrix $\mathbf{\Phi}_{NN}(f)$ to adjust to the current environment. Here we consider the initialization only.

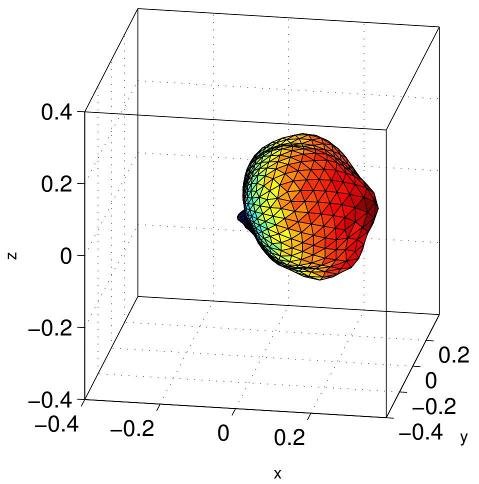

**Fig. 1**. Normalized measured directivity pattern for the rightmost microphone at 500 Hz. The pattern is approximately cardioid.

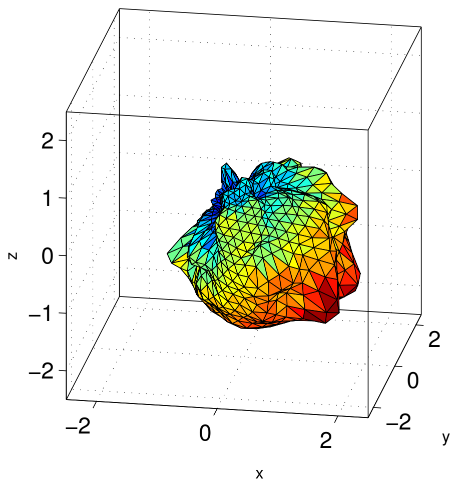

**Fig. 2**. Normalized measured directivity pattern for the rightmost microphone at 5 kHz. The influence of the enclosure leads to a more complicated and downward-facing pattern that is dissimilar to the standard cardioid model.

## 3. EXPERIMENTATION

### 3.1. Experimental Setup

The microphone array employed in Microsoft Kinect for Windows was used as an experimental test case. The array consists of four cardioid microphones in a nonuniform linear configuration, mounted in boots on the underside of a plastic enclosure. The assembly was designed to maximize the microphone directivity indices within the speech spectrum (200–7.2 kHz). Ten Microsoft Kinect for Windows arrays were obtained to account for manufacturing variations in the microphone capsules. One device was used to train the beamformer design and was excluded from the remaining test set.

The training device was placed in an anechoic chamber and aligned to face along the positive $x$-axis. The microphone directivity patterns were measured on an $11.25°$ equiangle grid by supervised estimation of the transfer function between a measurement loudspeaker and the array. Prior to conducting the experiment, the transfer function of the measurement loudspeaker was measured and equalized to reduce its influence upon the measurements. The optimization problem in (15) was then solved for three scenarios: (a) in 2D (azimuth only) using a standard cardioid model, (b) 3D (azimuth and elevation) cardioid model and (c) 3D measured model. A practical modification was made to the distortionless constraint in (15) so that $\mathbf{W}_0^T(f')\mathbf{D}_0(f') = 1$ for $200 \leq f' \leq 7500$ and 0 elsewhere. Further details on the 2D implementation can be found in [6]. In all cases, the spectrum of the isotropic ambient noise spectrum $N_0(f)$ and instrument noise spectrum $N_{\mathrm{I}}(f)$ were estimated from a corpus of average real-world noise recordings. The best-performing single microphone output was used as an additional reference.

A speech test set was created consisting of 6 clean sentence pairs spoken by 2 males, 2 females, and 2 children. The sentences were produced in a real noisy living room environment of approximately 2.8×5.6 m using a mouth simulator placed at 10 locations relative to the microphone array: 4 at range 1 m, 2 at 2 m, 2 at 3 m, and 2 at 4 m. Each sentence was produced at 65 dBSPL at 1 m to simulate typical talking levels. The presented results were averaged over all devices.

The processed speech quality was estimated in each case using ITU-T P.862.2 (PESQ) [8]. Automatic speech recognition (ASR) was performed using the Microsoft Speech Platform v.11.0[1] using the Kinect trained acoustic model from the Kinect Development Kit (KDK)[2]. The word error rate (WER) and sentence error rate (SER) were reported. As an additional measure of performance, the directivity index (DI) measures a beamformer's ability to suppress energy arriving from directions outside the look direction [6]:

$$\mathrm{DI}(f, \Omega_0) = 10 \log_{10} \left( \frac{P(f, \Omega_0)}{\int_{\Omega \in S^2} P(f, \Omega) \mathrm{d}\Omega} \right), \tag{18}$$

where $P(f, \Omega) = |B(f, \Omega)|^2$.

### 3.2. Discussion

The measured directivity patterns for the rightmost microphone are shown in Figs. 1 and 2 at 500 Hz and 5 kHz respectively. The directivity pattern at 500 Hz is similar to the expected cardioid pattern. At 5 kHz, the response is tilted downwards relative to $(0, 0)$ as a consequence of the underside mounting, rendering the cardioid model a poor approximation to the measured data. The microphone gain was also

[1] http://www.microsoft.com/download/en/details.aspx?id=27225
[2] http://www.microsoft.com/en-us/kinectforwindows/develop/

|  | PESQ | WER (%) | SER (%) |
|---|---|---|---|
| Best Mic | 2.13 | 18.47 | 31.67 |
| 2D Model | 2.62 | 9.67 | 15.00 |
| 3D Model | 2.64 | 9.79 | 15.00 |
| 3D Meas. | **2.66** | **4.92** | **9.17** |

**Table 1**. Microphone array performance metrics.

significantly higher at 5 kHz than at 500 Hz, causing the difference in scale between Figs. 1 and 2. The results are reported for the regularization parameter $\kappa$ that maximizes the PESQ score. A $\sim 0.5$ improvement in PESQ is achieved comparing the best single microphone with the beamformer output and appear largely invariant to the type of beamformer employed. Word and sentence error rates are similar for both 2D and 3D models. However there is a significant relative reduction in WER of approximately 50% (10 percentage points) comparing the 3D measured to the 3D model and 70% (13 percentage points) compared with the best microphone.

The plots in Figure 3 show the directivity indices, as a function of frequency, of the best-performing single microphone ($\circ$), beamformer using 3D models ($\Box$) and beamformer using 3D measured data ($\times$) for a farfield source located at $\Omega_0 = (0, 0)$. The DI results of the 2D model beamformer were near-identical to the 3D model beamformer and have not been plotted. The analytically-derived DI for a cardioid microphone is 4.8 dB [6] as confirmed by the measured data ($\circ$) in the frequency range ~500 Hz–3 kHz. Above 3 kHz, the DI reduces due to the downward tilt of the main lobe relative to $(0, 0)$ as seen in Figure 2. The beamformer designed with the 3D cardioid model provides a 2–4 dB improvement in DI compared with the best microphone in the range 1 kHz–5.5 kHz as reflected in the improved results in Table 1. Above 6 kHz, the downward rotation of the microphone main lobes has a compound effect on the 3D model beamformer's DI, reducing it to below that of the best microphone. In the case of the 3D measured beamformer, a 6 dB improvement in DI relative to the best microphone is observed throughout the range 2 kHz - 7.2 kHz. All three cases converge to the same DI at 500 Hz, at which point the wavelength of the incident wave becomes comparable to the size of the array aperture.

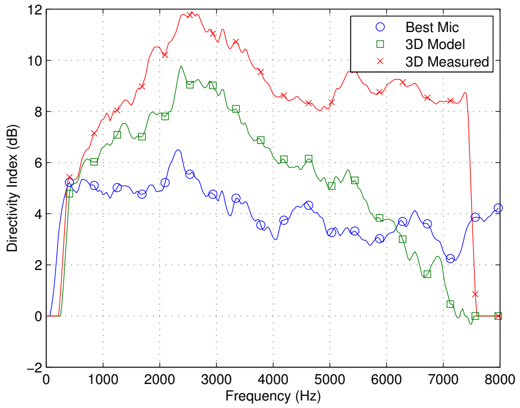

**Fig. 3**. Directivity indices $\mathrm{DI}(f, \Omega_0)$ as a function on frequency for best microphone ($\circ$), beamformer using 3D models ($\Box$) and beamformer using 3D measured data ($\times$).

## 4. CONCLUSIONS

A generalized solution for the MVDR beamforer has been proposed that exploits measured microphone directivity patterns as a function of azimuth, elevation and frequency. The additional information provided by such measurements allows more realistic design for those cases where the true directivity pattern deviates from standard microphone models. An illustrative example with the 4-element Microsoft Kinect for Windows array reveals that 2D models and 3D models yield beamformers with similar performance. Significant performance gains can however be achieved by designing the beamformer weights using measured data, reducing relative ASR word error rates on the test corpus by over 70% and improving directivity indices by 6 dB compared with the best microphone.

## 5. REFERENCES

[1] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Springer-Verlag, Berlin, Germany, 2001.

[2] H. Cox, R. M. Zeskind, and T. Kooij, "Practical supergain," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, pp. 393–398, June 1986.

[3] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[4] H. L. van Trees, *Optimum Array Processing*, Detection, Estimation and Modulation Theory. Wiley, 2002.

[5] I. Tashev and H. S. Malvar, "A new beamformer design algorithm for microphone arrays," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2005, vol. 3, pp. 101–104.

[6] I. Tashev, *Sound Capture and Processing: Practical Approaches*, Wiley, 2009.

[7] G. W Elko, "Superdirective microphone arrays," in *Acoustic Signal Processing for Telecommunications*, S. Gay and J. Benesty, Eds., chapter 10, pp. 181–237. Kluwer Academic, 2000.

[8] ITU-T P.862.2, "Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs," Nov. 2005.