Beamformer Design Using Measured Microphone Directivity Patterns: Robustness to Modelling Error

Mark R. P. Thomas, Jens Ahrens and Ivan J. Tashev
Microsoft Research, One Microsoft Way, Redmond WA 98052, USA
E-mail: {markth, jeahrens, ivantash}@microsoft.com

***Abstract*—The design process for time-invariant acoustic beamformers often assumes 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 investigate those cases in which one or more of these assumptions do not hold, considering a Minimum Variance Distortionless Response (MVDR)-based solution that is optimized using measured directivity patterns as a function of azimuth, elevation and frequency. Robustness to modelling error is controlled by a regularization parameter that produces a suboptimal but more robust solution. 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. Speech recognition and PESQ results are used as evaluation criteria with a noisy speech corpus, revealing empirically optimal regularization parameters for each case and up to a 70% relative improvement in word error rate comparing (a) and (c).**

I. Introduction

A microphone array is a device that spatially samples a soundfield within a finite aperture. A common application of the observed signals is a spatial filter called a beamformer [1] that exploits spatial diversity by combining the observations to enhance a desired signal, exceeding the performance that can be achieved using a single microphone. Beamformer design is subject to design constraints imposed by the limitations of the measurement apparatus and the nature of the desired/undesired signals. Adaptive (data-dependent) beamformers continuously optimize their design by estimating the ambient noise conditions [1]; in contrast, time-invariant beamformers use *a priori* assumptions about the operating environment. Superdirective beamforming [1] is an approach to time-invariant beamforming that 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 adaptive and time-invariant cases by estimating the expected noise cross-power density to minimize the output noise power [3], [4]. Several variants exist to address problems caused by sensor self-noise and sensor mismatch, for which white noise gain constraints [1], optimization with steering vector uncertainty sets [5], and optimization with the gain and phase error distributions [6] have been shown to successfully reduce such effects.

The design process for acoustic beamformers often assumes that the microphones are omnidirectional with a flat frequency response. In some real-world scenarios, such assumptions are not valid due to physical factors such as directional microphone responses and the acoustics of the mounting hardware. This increases the dimensionality of the design problem as the microphone response becomes a function of elevation and frequency in addition to azimuth. Beamformer design for arbitrary 2D directivity patterns and frequency responses was considered in [7], also accounting for ambient and instrumental noise spectra to yield a more realistic design. In this paper we provide additional insight into a generalized 3D solution proposed in [8] and investigate the performance by comparing designs based upon measured 3D directivity patterns and standard microphone models. This imposes no additional computational overhead at runtime as the design modifies the steering weights only. The incorporation of a regularization parameter improves robustness by preventing overfitting to the training data, providing a tradeoff between the optimal solution and the worse performing but more robust delay-and-sum beamformer.

The remainder of this paper is organized as follows. The beamformer problem is formulated in Section II and a solution is proposed based upon the MVDR criterion. In Section III, beamformers are designed for a 4-channel microphone array and speech recognition error rates are presented as a function of regularization parameter for designs based upon 2D microphone models, 3D microphone models and 3D measurements. Concluding remarks are given in Section IV.

II. Optimal Beamforming

A. 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. The 3D directivity pattern for an impinging wave from direction $\Omega = (\theta, \phi)$ to microphone m is $U_m(f, \Omega)$, where f denotes frequency (Hz), and $\theta = [-\pi/2, \pi/2]$ and $\phi = [0, 2\pi)$ are elevation and azimuth angles respectively. The midpoint of the array is the origin of the coordinate system. Let $S_0(f)$ be a farfield source located at angle Ω_0 in the frequency domain. The response of the array is

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

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

whose elements are

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

The term τ_m is the delay of the incoming wavefront at the mth sensor relative to the array centre. 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 Ω,

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

Given observations $\mathbf{X}(f)$, the output of a generalized filter-and-sum beamformer is [9],

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

where $\mathbf{W}_0^T(f)$ is an $M \times 1$ vector of complex weights computed to steer the beam in look direction Ω_0. The resulting directivity pattern is a weighted sum of the capture vector elements at angle Ω,

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

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

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

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

B. Calculation of Steering Weights

The design approach we employ is based upon the frequency domain minimum variance distortionless response (MVDR) beamformer [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 $Y(f) = S(f)$. In noisy conditions we aim to minimize the expected noise variance. Combining (1) and (4) we obtain a new expression for the beamformer output,

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

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

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

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

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

The frequency dependence of $\Phi_{ij}(f)$ has been dropped for simplicity. 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 [10]

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

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

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

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

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

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, usually by imposing models or using measured data. Such a design is however sensitive to instrumental noise, particularly in the lower end of the frequency spectrum [11]. Without appropriate modification of the design criteria, the ambient noise that is suppressed can be replaced by the amplified instrumental noise. An additional term is therefore added to $\mathbf{\Phi}_{NN}(f)$ to improve robustness [2]:

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

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 $N_{\mathrm{I}}(f)$, κ is a regularization parameter and $\mathbf{I}$ is the $M \times M$ identity matrix. In practice this lowers the directivity index but increases total noise suppression. The design procedure is summarized as a constrained optimization problem:

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

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]

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

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

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

The implementation of this algorithm 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 adapt to the current environment. Here we consider the initialization only.

III. Experimentation

A. Experimental Setup

The microphone array employed in Microsoft Kinect for Windows was used as an experimental test case. It consists of four cardioid microphones mounted on the underside of a plastic enclosure in a nonuniform linear configuration that is acoustically designed to maximize the microphone directivity

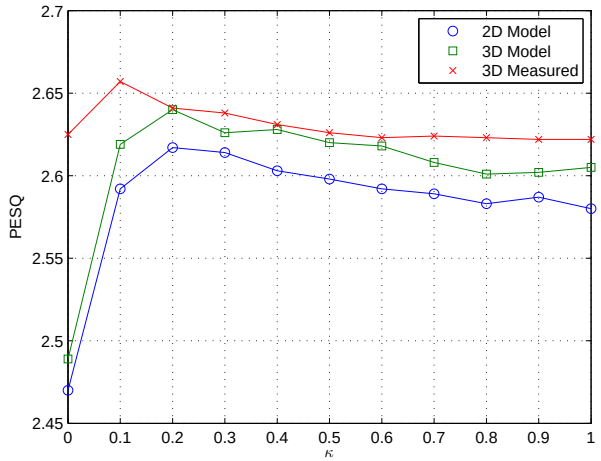

Fig. 1. PESQ scores as a function of regularization parameter κ.

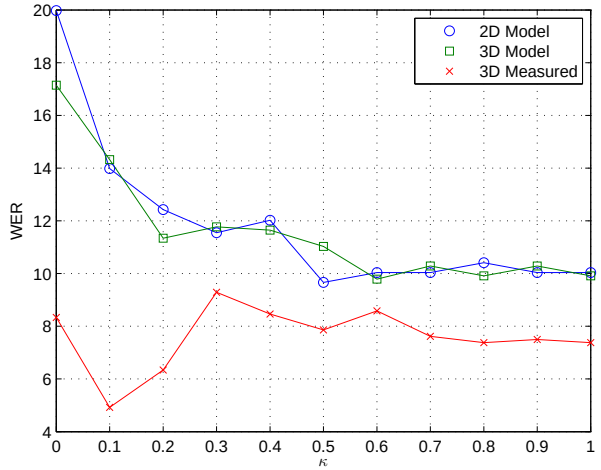

Fig. 2. Word error rate as a function of regularization parameter κ.

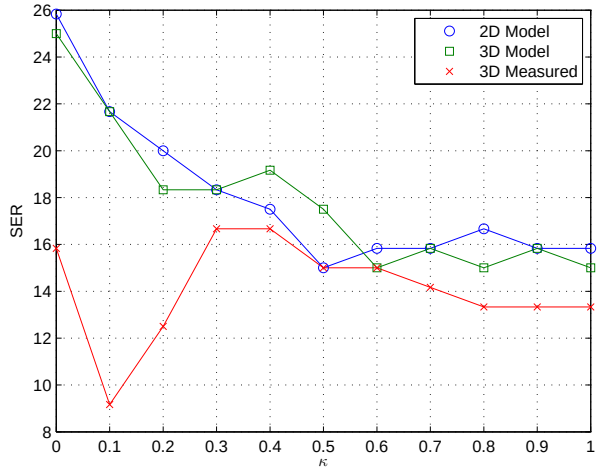

Fig. 3. Sentence error rate as a function of regularization parameter κ.

	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 I
BEAMFORMER PERFORMANCE METRICS.

indices within the speech spectrum (200 Hz–7.2 kHz). Ten Microsoft Kinect for Windows arrays were obtained, one of which was used to train the beamformer design; the remainder were used as a test set that incoporates manufacturing variations in the microphone capsules.

The training device was placed in an anechoic chamber and aligned to face along the positive x-axis. An array of measurement loudspeakers was moved on a circular trajectory to obtain a supervised estimate of the free-field microphone directivity patterns on an 11.25° equiangle grid. The transfer function of the measurement loudspeaker was measured and equalized to reduce its influence upon the measurements. The design problem (15) was then solved for three scenarios: (a) 2D cardioid model (azimuth only), (b) 3D cardioid model (azimuth and elevation) and (c) 3D measured model. The solutions $\mathbf{W}_0$ were calculated for regularization parameters κ in the range $0 \leq \kappa \leq 1$ in steps of 0.1. 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. In all cases, the ambient noise spectrum $N_0(f)$ was assumed to be isotropic. The instrument noise spectrum $N_\mathrm{I}(f)$ was measured with a microphone of the type used in the Kinect array. The ambient noise $N_0(f)$ noise spectrum was estimated from a corpus of noise recordings compiled in living room environments representative of Kinect's target locations. Further details on the 2D implementation can be found in [9].

A speech test corpus was created consisting of 2 males, 2 females, and 2 children speaking 6 short sentences. 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 dB SPL at 1 m to simulate typical talking levels. The best-performing single microphone was used as an additional reference.

The processed speech quality was estimated using ITU-T P.862 (PESQ) [12]. Automatic speech recognition (ASR) was performed using the Microsoft Speech Platform v.11.0[1] using the trained acoustic model from the Kinect Development Kit (KDK)[2], with which word error rate (WER) and sentence error rate (SER) were calculated. The results were averaged over all devices.

B. Discussion

Figs. 1–3 show the PESQ score, WER and SER as a function of regularization parameter κ. They reveal that reg-

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

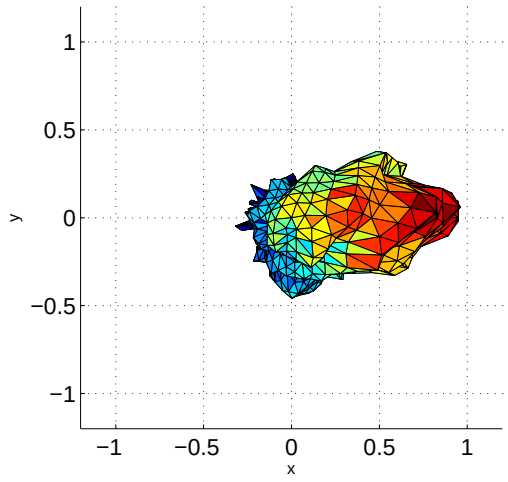

Fig. 4. Measured beamformer directivity pattern viewed down the z-axis at 1 kHz using weights derived from 2D microphone models, $\kappa = 0.5$.

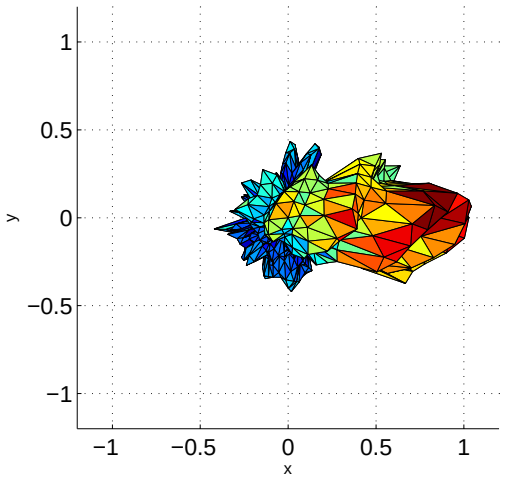

Fig. 5. Measured beamformer directivity pattern viewed down the z-axis at 1 kHz using weights derived from 3D measured data, $\kappa = 0.1$.

ularization improves the performance of all three solutions by avoiding over-fitting to the training data. The optimal 3D measured performance for all three metrics is near the relatively low value of $\kappa \simeq 0.1$; in contrast, the PESQ score in the 2D and 3D model solutions is maximized at $\kappa \simeq 0.2$ but WER and SER benefit from much higher values. The variance of the PESQ results is however very small therefore WER/SER are of greater interest. The requirement for higher κ for the 2D/3D model solutions suggests that these designs are less suited to the test corpus than the solution based upon 3D measurements.

The average results for the test corpus are shown in Table I using the empirical optimum regularization parameters for WER in Fig. 2. The improvement in PESQ of about 0.5 points over a single microphone is largely invariant to the type of beamformer. 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% absolute) comparing the 3D measured to the 3D model and 70% (13% absolute) compared with the best microphone. The measured beamformer directivity patterns using weights derived from the 2D microphone model and the 3D measured microphones are shown looking down the z-axis at 1 kHz in Figs. 4 and 5 respectively, the latter displaying a narrower main lobe.

IV. Conclusions

A generalized solution for the MVDR beamforer has been investigated that exploits measured microphone directivity patterns as a function of azimuth, elevation and frequency. The use of measured directivity patterns allows more realistic design for those cases where the true directivity pattern deviates from standard microphone models. It incorporates a regularization parameter that provides robustness to mismatch between training and test data caused by manufacturing variations between devices. Experimental results with the 4-element Microsoft Kinect for Windows array reveals that significant performance gains can 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 single microphone. The best 3D measured data design results are achieved using a lower regularization parameter than the 2D model, showing that design with measurements from a single training device are applicable to a much broader test corpus.

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] R. Lorenz and S. Boyd, "Robust minimum variance beamforming," *IEEE Trans. Signal Process.*, vol. 53, no. 5, pp. 1684–1696, May 2005.

[6] S. Doclo and M. Moonen, "Design of broadband beamformers robust against gain and phase errors in the microphone array characteristics," *IEEE Trans. Signal Process.*, vol. 51, no. 10, pp. 2511–2526, Oct. 2003.

[7] 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.

[8] M. R. P. Thomas, J. Ahrens, and I. Tashev, "Optimal 3D beamforming using measured microphone directivity patterns," in *Proc. Intl. Workshop Acoustic Signal Enhancement (IWAENC)*, Aachen, Germany, Sept. 2012.

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

[10] 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.

[11] I. Tashev and D. Allred, "Reverberation reduction for improved speech recognition," in *Proc. Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, Piscataway, NJ, USA, Mar. 2005.

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