Microphone Array Processing for Distant Speech Recognition: Spherical Arrays

John McDonough*, Kenichi Kumatani†, Bhiksha Raj‡
* Carnegie Mellon University, Voci Technologies, Inc., Pittsburgh, PA, USA
E-mail: johnmcd@cs.cmu.edu
† Disney Research, Pittsburgh, USA
E-mail: kenichi.kumatani@disneyresearch.com
‡ Carnegie Mellon University, Pittsburgh, PA, USA
E-mail: bhiksha@cs.cmu.edu

***Abstract*—Distant speech recognition (DSR) holds out the promise of the most natural human computer interface because it enables man-machine interactions through speech, without the necessity of donning intrusive body- or head-mounted microphones. With the advent of the Microsoft Kinect, the application of non-uniform linear arrays to the DSR problem has become commonplace. Performance analysis of such arrays is well-represented in the literature. Recently, spherical arrays have become the subject of intense research interest in the acoustic array processing community. Such arrays have heretofore been analyzed solely with theoretical metrics under idealized conditions. In this work, we analyze such arrays under realistic conditions. Moreover, we compare a linear array with 64-channel arrays and a total length of 126 cm to a spherical array with 32 channels and a radius of 4.2 cm; we found that these provided word error rates of 9.3% and 10.2%, respectively, on a DSR task. For a speaker positioned at an oblique angle with respect to the linear array, we recorded error rates of 12.8% and 9.7%, respectively, for the linear and spherical arrays. The compact size and outstanding performance of the spherical array recommends itself well to space-limited and mobile applications such as home-gaming consoles and humanoid robots.**

I. Introduction

When the signals from the individual sensors of a microphone array with a known geometry are suitably combined, the array functions as a spatial filter capable of suppressing noise, reverberation, and competing speech. Such *beamforming* techniques have received a great deal of attention within the acoustic array processing community in the recent past [1], [2], [3], [4], [5], [6], [7]. With the advent of the Microsoft Kinect, the application of non-uniform linear arrays to the *distant speech recognition* (DSR) problem has become commonplace. Performance analysis of such arrays is well-represented in the literature [8]. Recently, spherical arrays have become the subject of intense research interest in the acoustic array processing community [9]. Such arrays have heretofore been analyzed solely with theoretical metrics under idealized conditions [10], [11]; e.g., assuming an ideal array with a continous pressure sensitive surface. The effects of *discretization*, whereby the continuous pressure-sensitive surface must be replaced with a finite number of discrete sensors, are typically ignored. In this contribution, we analyze the effects of such discretization using the theoretical metrics of *array gain*, *white noise gain*, and *directivity index*; these metrics are defined and discussed in Section III. More importantly, we compare a conventional linear array with a spherical array in terms of *word error rate*, the preferred metric in the DSR literature. To the knowledge of the present authors, such a study has never been undertaken in the literature.

The present contribution complements Kumatani *et al* [12], also appearing in these proceedings, with some amount of unavoidable overlap. Much of the material appearing here is based on Kumatani *et al* [13] as well as the recent book chapter McDonough and Kumatani [14], and is intended to serve as an introduction to these works.

The remainder of this article is organized as follows. In Section II we consider the conventional beamforming techniques, including delay-and-sum, minimum variance distortionless response, and superdirective designs. A review of the performance metrics used in conventional beamforming is discussed in Section III. The fundamentals of the analysis of spherical arrays is presented in Section IV, and Section V describes how the beamformer designs for conventional arrays can be adapted for spherical arrays. The principal results of this section are presented in Section VII, wherein a conventional linear array is compared to its spherical counterpart in terms of the conventional performance criteria from Section III, as well as in terms of the all important word error rate, the metric of choice for DSR.

II. Conventional Beamforming Techniques

Here we introduce the conventional beamforming designs for conventional arrays; much of this material necessarily overlaps with Kumatani *et al* [12]. The presentation of these designs will lead naturally to the discussion of their counterparts for spherical arrays.

In the case of the spherical wavefront depicted in Figure 1a, let us define the *propagation delay* as $\tau_s \triangleq D_s/c$. In the far-field case shown in Figure 1b, let us define the *wavenumber* $\mathbf{k}$ as a vector perpendicular to the planar wavefront pointing in the direction of propagation with magnitude $\omega/c = 2\pi/\lambda$. Then, the propagation delay with respect to the origin of the coordinate system for microphone s is determined through $\omega\tau_s = \mathbf{k}^T\mathbf{m}_s$.

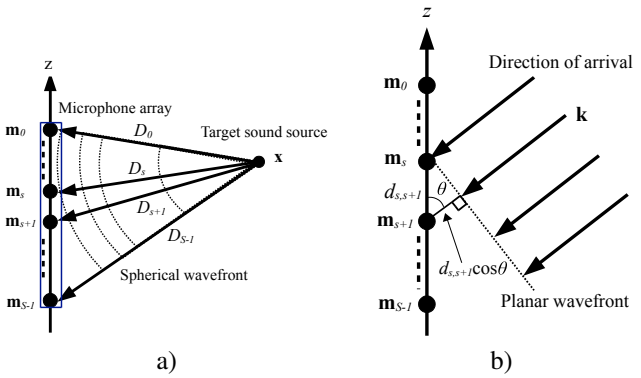


Fig. 1. Propagation of a) the spherical wave and b) plane wave.

A. Delay-and-Sum Beamformer

The simplest model of wave propagation assumes that a signal $f(t)$ at t, carried on a plane wave, reaches all sensors in an array, but not at the same time. Hence, let us form the vector

$$\mathbf{f}(t) = \begin{bmatrix} f(t-\tau_0) & f(t-\tau_1) & \cdots & f(t-\tau_{S-1}) \end{bmatrix}^T$$

of the time delayed signals reaching each sensor s, where S is the total number of sensors. In the frequency domain, the comparable vector of *phase-delayed* signals is $\mathbf{F}(\omega) = F(\omega)\mathbf{v}(\mathbf{k},\omega)$ where $F(\omega)$ is the transform of $f(t)$ and

$$\mathbf{v}(\mathbf{k},\omega) \triangleq \begin{bmatrix} e^{-i\omega\tau_0} & e^{-i\omega\tau_1} & \cdots & e^{-i\omega\tau_{S-1}} \end{bmatrix}^T \qquad (1)$$

is the *array manifold vector*, which is manifestly a vector of phase delays for a plane wave with wavenumber $\mathbf{k}$. To a first order, the array manifold vector is a complete description of the interaction of a propagating wave and an array of sensors. Note that the notation $\mathbf{v}(\mathbf{k},\omega)$ is actually redundant in that the magnitude of $\mathbf{k}$ is ω/c.

If $\mathbf{X}(\omega)$ denotes the vector of frequency domain signals for all sensors, the so-called *snapshot vector*, and $Y(\omega)$ the frequency domain output of the array, then the operation of a beamformer can be represented as

$$Y(\omega) = \mathbf{w}^H(\omega)\,\mathbf{X}(\omega), \qquad (2)$$

where $\mathbf{w}(\omega)$ is a vector of frequency-dependent sensor weights. The differences between various beamformer designs are completely determined by the specification of the weight vector $\mathbf{w}(\omega)$. The simplest beamforming algorithm, the *delay-and-sum* (DS) beamformer, time aligns the signals for a plane wave arriving from the look direction by setting

$$\mathbf{w}_{\text{DS}} \triangleq \mathbf{v}(\mathbf{k},\omega)/S. \qquad (3)$$

Substituting $\mathbf{X}(\omega) = \mathbf{F}(\omega) = F(\omega)\mathbf{v}(\mathbf{k},\omega)$ into (11) provides

$$Y(\omega) = \mathbf{w}_{\text{DS}}^H(\omega)\,\mathbf{v}(\mathbf{k},\omega)\,F(\omega) = F(\omega);$$

i.e., the output of the array is equivalent to the original signal in the absence of any interference or distortion. In general, this will be true for any weight vector achieving

$$\mathbf{w}^H(\omega)\,\mathbf{v}(\mathbf{k},\omega) = 1. \qquad (4)$$

Hereafter we will say that any weight vector $\mathbf{w}(\omega)$ achieving (4) satisfies the *distortionless constraint*, which implies that any wave impinging from the look direction is neither amplified nor attenuated.

B. Minimum Variance Distortionless Response Beamformer

To improve upon noise suppression performance provided by the DS beamformer, it is possible to adaptively suppress spatially-correlated noise and interference $\mathbf{N}(\omega)$, which can be achieved by adjusting the weights of a beamformer so as to minimize the variance of the noise and interference at the output subject to the distortionless constraint (4). More concretely, we seek $\mathbf{w}(\omega)$ achieving

$$\text{argmin}_{\mathbf{W}}\ \mathbf{w}^H(\omega)\,\mathbf{\Sigma}_{\mathbf{N}}(\omega)\,\mathbf{w}(\omega), \qquad (5)$$

subject to (4), where $\mathbf{\Sigma}_{\mathbf{N}} \triangleq \mathcal{E}\{\mathbf{N}(\omega)\mathbf{N}^H(\omega)\}$ and $\mathcal{E}\{\cdot\}$ is the expectation operator. In practice, $\mathbf{\Sigma}_{\mathbf{N}}$ is computed by averaging or recursively updates the noise covariance matrix [8, §7]. The weight vectors obtained under these conditions correspond to the *minimum variance distortionless response* (MVDR) beamformer, which has the well-known solution [2, §13.3.1]

$$\mathbf{w}_{\text{MVDR}}^H(\omega) = \frac{\mathbf{v}^H(\mathbf{k},\omega)\,\mathbf{\Sigma}_{\mathbf{N}}^{-1}(\omega)}{\mathbf{v}^H(\mathbf{k},\omega)\,\mathbf{\Sigma}_{\mathbf{N}}^{-1}(\omega)\,\mathbf{v}(\mathbf{k},\omega)}. \qquad (6)$$

If $\mathbf{N}(\omega)$ consists of a single plane interferer with wavenumber $\mathbf{k}_{\text{I}}$ and spectrum $N(\omega)$, then $\mathbf{N}(\omega) = N(\omega)\mathbf{v}(\mathbf{k}_{\text{I}})$ and $\mathbf{\Sigma}_{\mathbf{N}}(\omega) = \Sigma_N(\omega)\mathbf{v}(\mathbf{k}_{\text{I}})\mathbf{v}^H(\mathbf{k}_{\text{I}})$, where $\Sigma_N(\omega) = \mathcal{E}\{|N(\omega)|^2\}$.

Depending on the acoustic environment, adapting the sensor weights $\mathbf{w}(\omega)$ to suppress discrete sources of interference can lead to excessively large sidelobes, resulting in poor system robustness. A simple technique for avoiding this is to impose a quadratic constraint $\|\mathbf{w}\|^2 \leq \gamma$, for some $\gamma > 0$, in addition to the distortionless constraint (4), when estimating the sensor weights. The MVDR solution will then take the form [2, §13.3.7]

$$\mathbf{w}_{\text{DL}}^H = \frac{\mathbf{v}^H\left(\mathbf{\Sigma}_{\mathbf{N}} + \sigma_{\text{d}}^2\mathbf{I}\right)^{-1}}{\mathbf{v}^H\left(\mathbf{\Sigma}_{\mathbf{N}} + \sigma_{\text{d}}^2\mathbf{I}\right)^{-1}\mathbf{v}}, \qquad (7)$$

which is referred to as *diagonal loading* where σ_{d}^2 is the loading level; the dependence on ω in (7) has been suppressed for convenience. While (7) is straightforward to implement, there is no direct relationship between γ and σ_{d}^2; hence the latter is typically set either based on experimentation or through an iterative procedure. Increasing σ_{d}^2 decreases $\|\mathbf{w}_{\text{DL}}\|$, which implies that the *white noise gain* (WNG) also increases [15]; WNG is a measure of the robustness of the system to steering errors, as well as errors in sensor placement and response characteristics.

C. Super-Directive Beamformer

A theoretical model of diffuse noise that works well in practice is the spherically isotropic field, wherein spatially separated microphones receive equal energy and random phase noise signals from all directions simultaneously [16, §4]. The MVDR beamformer with the diffuse noise model is called the *super-directive beamformer* [2, §13.3.4]. The super-directive beamforming design is obtained by replacing the

noise covariance matrix $\mathbf{\Sigma_N}(\omega)$ with the coherence matrix $\mathbf{\Gamma}(\omega)$ whose (m,n)-th component is given by

$$\Gamma_{m,n}(\omega) = \text{sinc}\left(\frac{\omega d_{m,n}}{c}\right), \tag{8}$$

where $d_{m,n}$ is the distance between the mth and nth elements of the array, and sinc $x \triangleq \sin x/x$. Notice that the weight of the super-directive beamformer is determined solely based on the distance between the sensors $d_{m,n}$ and is thus data-independent. In the most general case, the acoustic environment will consist both of diffuse noise as well as one or more sources of discrete interference, such as in

$$\mathbf{\Sigma_N}(\omega) = \Sigma_N(\omega)\mathbf{v}(\mathbf{k}_\text{I})\mathbf{v}^H(\mathbf{k}_\text{I}) + \sigma_\text{SI}^2\mathbf{\Gamma}(\omega), \tag{9}$$

where σ_SI^2 is the power spectral density of the diffuse noise.

D. Minimum Mean-Square Error Beamformer

The MVDR beamformer is of particular interest because it forms the preprocessing component of two other important beamforming structures. Firstly, the MVDR beamformer followed by a suitable post-filter yields the *maximum signal-to-noise ratio* beamformer [8, §6.2.3]. Secondly, and more importantly, by placing a Wiener filter [17, §2.2] on the output of the MVDR beamformer, the *minimum mean-square error* (MMSE) beamformer is obtained [8, §6.2.2]. Such *post-filters* are important because it has been shown that they can yield significant reductions in error rate [5], [18], [19], [20]. Of the several post-filtering methods proposed in the literature [21], the *Zelinski post-filtering* [22] technique is arguably the simplest practical implementation of a Wiener filter. Wiener filters in their pure form are unrealizable because they assume that the spectrum of the desired signal is available. The Zelinski post-filtering method uses the auto- and cross-power spectra of the multi-channel input signals to estimate the target signal and noise power spectra effectively under the assumption of zero cross-correlation between the noises at different sensors.

The MVDR beamformer and its variants can effectively suppress sources of interference. They can also potentially cancel the target signal, however, in cases wherein signals correlated with the target signal arrive from directions other than the look direction. This is precisely what happens in all real acoustic environments due to reflections from hard surfaces such as tables, walls and floors. A brief overview of techniques for preventing signal cancellation can be found in [23].

III. Beamforming Performance Criteria

Before continuing our discussion of adaptive array processing algorithms, we introduce three measures of beamforming performance, namely, the *array gain*, *white noise gain*, and the *directivity index*. These criteria will prove useful in our performance comparisons of conventional, linear and spherical arrays in Section V.

A. Array Gain

The array gain is defined as the ratio of the *signal-to-noise* (SNR) ratio at the output of the beamformer to the SNR at the input of a single channel of the array. Hence, array gain is a useful measure of how much a particular acoustic array processing algorithm enhances the desired signal. In this section, we formalize the concept of the array gain, and calculate it for both the delay-and-sum and MVDR beamformers given in (3) and (6), respectively.

Let us assume that the component of the desired signal reaching each component of a sensor array is $F(\omega)$ and the component of the noise and interference reaching each sensor is $N(\omega)$. This implies that the SNR at the input of the array can be expressed as

$$\text{SNR}_\text{in}(\omega) \triangleq \frac{\Sigma_F(\omega)}{\Sigma_N(\omega)}, \tag{10}$$

where $\Sigma_F(\omega) \triangleq \mathcal{E}\{|F(\omega)|^2\}$ and $\Sigma_N(\omega) \triangleq \mathcal{E}\{|N(\omega)|^2\}$. Then for the vector of beamforming weights $\mathbf{w}^H(\omega)$, the output of the array is given by

$$Y(\omega) = \mathbf{w}^H(\omega)\,\mathbf{X}(\omega) = Y_F(\omega) + Y_N(\omega), \tag{11}$$

where $Y_F(\omega) \triangleq \mathbf{w}^H(\omega)\,\mathbf{F}(\omega)$ and $Y_N(\omega) \triangleq \mathbf{w}^H(\omega)\,\mathbf{N}(\omega)$ are, respectively, the signal and noise components in the output of the beamformer. Let us define the *spatial spectral covariance matrices*

$$\mathbf{\Sigma_F}(\omega) \triangleq \mathcal{E}\{\mathbf{F}(\omega)\mathbf{F}^H(\omega)\},$$
$$\mathbf{\Sigma_N}(\omega) \triangleq \mathcal{E}\{\mathbf{N}(\omega)\mathbf{N}^H(\omega)\}.$$

Then, upon assuming the $F(\omega)$ and $N(\omega)$ are statistically independent, the variance of the output of the beamformer can be calculated according to

$$\Sigma_Y(\omega) = \mathcal{E}\{|Y(\omega)|^2\} = \Sigma_{Y_F}(\omega) + \Sigma_{Y_N}(\omega), \tag{12}$$

where

$$\Sigma_{Y_F}(\omega) \triangleq \mathbf{w}^H(\omega)\,\mathbf{\Sigma_F}(\omega)\,\mathbf{w}(\omega) \tag{13}$$

is the variance of the signal component of the beamformer output, and

$$\Sigma_{Y_N}(\omega) \triangleq \mathbf{w}^H(\omega)\,\mathbf{\Sigma_N}(\omega)\,\mathbf{w}(\omega) \tag{14}$$

is the variance of the noise component. The spatial spectral matrix $\mathbf{F}(\omega)$ of the desired signal can be written as

$$\mathbf{\Sigma_F}(\omega) = \Sigma_F(\omega)\,\mathbf{v_k}(\mathbf{k}_\text{s})\,\mathbf{v_k}^H(\mathbf{k}_\text{s}), \tag{15}$$

where $\mathbf{k}_\text{s}$ is the wavenumber of the desired source, and $\mathbf{v_k}(\mathbf{k}_\text{s})$ is the array manifold vector (1). Substituting (15) into (13), we can calculate the variance of the output signal spectrum as

$$\Sigma_{Y_F}(\omega) = \mathbf{w}^H(\omega)\,\mathbf{v_k}(\mathbf{k}_\text{s})\,\Sigma_F(\omega)\,\mathbf{v_k}^H(\mathbf{k}_\text{s})\,\mathbf{w}(\omega). \tag{16}$$

If we now assume that $\mathbf{w}(\omega)$ satisfies the distortionless constraint (4), then (16) reduces to

$$\Sigma_{Y_F}(\omega) = \Sigma_F(\omega),$$

which holds for both the delay-and-sum and MVDR beamformers.

Substituting (3) into (14), it follows that the noise component present at the output of the DSB is given by

$$\Sigma_{Y_N}(\omega) = \frac{1}{S^2}\mathbf{v}_{\mathbf{k}}^H(\mathbf{k}_\mathrm{s})\,\mathbf{\Sigma}_{\mathbf{N}}(\omega)\,\mathbf{v}_{\mathbf{k}}(\mathbf{k}_\mathrm{s}) \tag{17}$$

$$= \frac{1}{S^2}\mathbf{v}_{\mathbf{k}}^H(\mathbf{k}_\mathrm{s})\rho_{\mathbf{N}}(\omega)\mathbf{v}_{\mathbf{k}}(\mathbf{k}_\mathrm{s})\Sigma_N(\omega), \tag{18}$$

where the *normalized spatial spectral matrix* $\rho_{\mathbf{N}}(\omega)$ is defined through the relation

$$\mathbf{\Sigma}_{\mathbf{N}}(\omega) \triangleq \Sigma_N(\omega)\,\rho_{\mathbf{N}}(\omega). \tag{19}$$

Hence, the SNR at the output of the beamformer is given by

$$\mathrm{SNR}_{\mathrm{out}}(\omega) \triangleq \frac{\Sigma_{Y_F}(\omega)}{\Sigma_{Y_N}(\omega)} = \frac{\Sigma_F(\omega)}{\mathbf{w}^H(\omega)\,\mathbf{\Sigma}_{\mathbf{N}}(\omega)\mathbf{w}(\omega)}. \tag{20}$$

Then based on (10) and (20), we can calculate the array gain of the DSB as

$$A_{\mathrm{dsb}}(\omega,\mathbf{k}_\mathrm{s}) \triangleq \frac{\Sigma_{Y_F}(\omega)}{\Sigma_{Y_N}(\omega)}\Big/\frac{\Sigma_F(\omega)}{\Sigma_N(\omega)} \tag{21}$$

$$= \frac{S^2}{\mathbf{v}_{\mathbf{k}}^H(\mathbf{k}_\mathrm{s})\,\rho_{\mathbf{N}}(\omega)\,\mathbf{v}_{\mathbf{k}}(\mathbf{k}_\mathrm{s})}. \tag{22}$$

Repeating the foregoing analysis for the MVDR beamformer (6), we arrive at

$$A_{\mathrm{mvdr}}(\omega,\mathbf{k}_\mathrm{s}) = \mathbf{v}_{\mathbf{k}}^H(\mathbf{k}_\mathrm{s})\,\rho_{\mathbf{N}}^{-1}(\omega)\,\mathbf{v}_{\mathbf{k}}(\mathbf{k}_\mathrm{s}). \tag{23}$$

If noise at all sensors are spatially uncorrelated, then $\rho_{\mathbf{N}}(\omega)$ is the identity matrix and the MVDR beamformer reduces to the DSB. From (22) and (23), it can be seen that in this case, the array again is

$$A_{\mathrm{mvdr}}(\omega,\mathbf{k}_\mathrm{s}) = A_{\mathrm{dsb}}(\omega,\mathbf{k}_\mathrm{s}) = S. \tag{24}$$

In all other cases,

$$A_{\mathrm{mvdr}}(\omega,\mathbf{k}_\mathrm{s}) > A_{\mathrm{dsb}}(\omega,\mathbf{k}_\mathrm{s}). \tag{25}$$

The MVDR beamformer is of particular interest because it comprises the preprocessing component of two other important beamforming structures. Firstly, the MVDR beamformer followed by a suitable post-filter yields the *maximum signal-to-noise ratio* beamformer [8, §6.2.3]. Secondly, and more importantly, by placing a Wiener filter [17, §2.2] on the output of the MVDR beamformer, the *minimum mean-square error* (MMSE) beamformer is obtained [8, §6.2.2]. Such *post-filters* are important because it has been shown that they can yield significant reductions in error rate [18], [24]. If only a single subband is considered, the MVDR beamformer without modification will uniformly provide the highest SNR, as indicated by (25), and hence the highest array gain; we will return to this point in Section V.

B. White Noise Gain

The *white noise gain* (WNG) is by definition [15]

$$G_{\mathrm{w}}(\omega) \triangleq \frac{\left|\mathbf{w}^H(\omega)\,\mathbf{v}(\mathbf{k}_\mathrm{s})\right|^2}{\mathbf{w}^H(\omega)\,\mathbf{w}(\omega)}. \tag{26}$$

The numerator of (26), which will be unity for any beamformer satisfying the distortionless constraint (4), represents the power of the desired signal at the output of the beamformer, while the denominator is equivalent to the array's sensitivity to self sensor noise. Gilbert and Morgan [25] explain that WNG also reflects the sensitivity of the array to random variations in its components, including the positions and response characteristics of its sensors. Hence, WNG is a useful measure of system robustness.

It can be shown that uniform weighting of the sensor outputs provides the highest WNG [8, §2.6.3]. Hence, we should expect the delay-and-sum beamformer to provide the highest WNG in all conditions; we will re-examine this assumption in Section V.

C. Directivity Index

We now describe our third beamforming performance metric. Let us begin by defining the *power pattern* as

$$P(\theta,\phi) \triangleq \left|B(\theta,\phi)\right|^2, \tag{27}$$

where $B(\theta,\phi)$ is the beampattern as a function of the spherical coordinates $\Omega \triangleq (\theta,\phi)$. Let $\Omega_0 \triangleq (\theta_0,\phi_0)$ denote the look direction. The *directivity* is typically defined in the traditional (i.e., non-acoustic) array processing literature as [8, §2.6.1]

$$D(\omega) \triangleq \frac{4\pi P(\theta_0,\phi_0)}{\displaystyle\int_{\Omega_{\mathrm{sph}}} P(\theta,\phi)\,d\Omega}, \tag{28}$$

where Ω_{sph} represents the surface of a sphere with differential area $d\Omega$; we will consider such spherical integrals in detail in Sections IV and V.

Assuming that the beamforming coefficients satisfy the distortionless constraint (4) implies $P(\Omega_0) = 1$ such that (28) can be simplified and expressed in decibels as the *directivity index*

$$\mathrm{DI} \triangleq -10\log_{10}\left[\frac{1}{4\pi}\int_\Omega P(\theta,\phi)\,d\Omega\right]$$

$$= -10\log_{10}\left[\frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi} P(\theta,\omega)\sin\theta\,d\theta\,d\phi\right]. \tag{29}$$

Note the critical difference between array gain and directivity index. While the former requires specific knowledge of the acoustic environment in which a given beamformer operates, the latter is the ratio of the sensitivity of the array in the look direction to that averaged over the surface of the sphere. Hence, the directivity index is independent of the acoustic environment once the beamforming weights have been specified.

In the acoustic array processing literature, directivity is more often defined as SNR in the presence of a spherically isotropic diffuse noise field with sensor covariance matrix defined in (8);

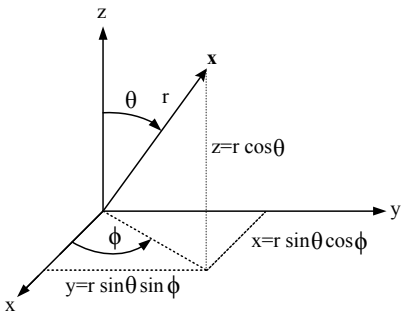


Fig. 2. Relationship between the Cartesian and spherical coordinate systems.

see [26]. Under this definition, the directivity index can be expressed as

$$\mathrm{DI} \triangleq -10\log_{10}\frac{\left|\mathbf{w}^H\mathbf{v}(\mathbf{k}_S)\right|^2}{\mathbf{w}^H\mathbf{\Gamma}_{SI}\mathbf{w}}. \tag{30}$$

The superdirective beamformer mentioned in Section II-C will uniformly provide the highest directivity index, although this may not be the case when the covariance matrix (8) is diagonally loaded to achieve greater robustness. We will return to this point in Section V.

IV. Spherical Microphone Arrays

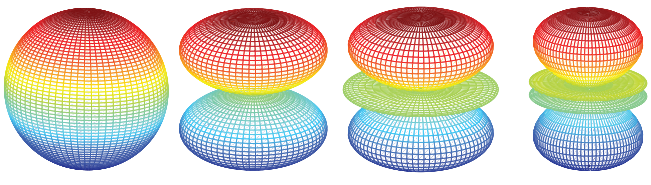
The advantage of spherical arrays is that they can be pointed at a desired speaker in any direction with equal effect; the shape of the beampattern is invariant to the look direction. The following sections provide a review of beamforming methods in the spherical harmonics domain. Thereafter we provide a comparison of spherical and linear arrays in terms of DSR performance.

In this section, we describe how beamforming is performed in the spherical harmonics domain. We will use the spherical coordinate system (r, θ, ϕ) shown in Figure 2 and denote the pair of *polar angle* θ and *azimuth* ϕ as $\Omega = (\theta, \phi)$.

Spherical Harmonics: Let us begin by defining the *spherical harmonic* of order n and degree m [9] as

$$Y_n^m(\Omega) \triangleq \sqrt{\frac{(2n+1)}{4\pi}\frac{(n-m)!}{(n+m)!}}P_n^m(\cos\theta)e^{im\phi}, \tag{31}$$

where $P_n^m(\cdot)$ denotes the associated Legendre function [27, §6.10.1]. Figure 3 shows the magnitude for the spherical harmonics, $Y_0 \triangleq Y_0^0$, $Y_1 \triangleq Y_1^0$, $Y_2 \triangleq Y_2^0$ and $Y_3 \triangleq Y_3^0$ in three-dimensional space. The spherical harmonics satisfy



a) b) c) d)

Fig. 3. Magnitude of spherical harmonics, a) Y_0, b) Y_1, c) Y_2 and d) Y_3.

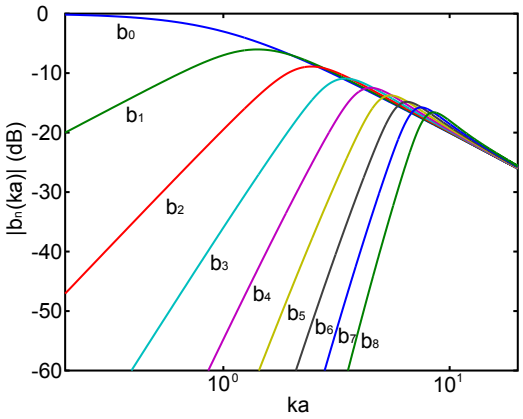


Fig. 4. Magnitude of the modal coefficients as a function of ka.

the *orthonormality condition* [9], [28],

$$\delta_{n,n'}\,\delta_{m,m'} = \int_\Omega Y_{n'}^{m'}(\Omega)\bar{Y}_n^m(\Omega)d\Omega \tag{32}$$

$$= \int_0^{2\pi}\int_0^{\pi} Y_{n'}^{m'}(\theta,\phi)\bar{Y}_n^m(\theta,\phi)\sin\theta d\theta\, d\phi, \tag{33}$$

where $\delta_{m,n}$ is the Kronecker delta function, and $\bar{Y}$ is the complex conjugate of Y.

Spherical Fourier Transform: In Section II, we defined the wavenumber as a vector perpendicular to the front of a plane wave of angular frequency ω pointing in the direction of propagation with a magnitude of ω/c. Now let us define the *wavenumber scalar* as $k = |\mathbf{k}| = \omega/c$; when no confusion can arise, we will also refer to k as simply the wavenumber. Let us assume that a plane wave of wavenumber k with unit power is impinging on a rigid sphere of radius a from direction $\Omega_0 = (\theta_0, \phi_0)$. The total complex sound pressure on the sphere surface at Ω_s can be expressed as

$$G(ka, \Omega_s, \Omega_0) = 4\pi\sum_{n=0}^{\infty} i^n b_n(ka)\sum_{m=-n}^{n}\bar{Y}_n^m(\Omega_0)Y_n^m(\Omega_s), \tag{34}$$

where the *modal coefficient* $b_n(ka)$ is defined as [9], [10]

$$b_n(ka) \triangleq j_n(ka) - \frac{j_n'(ka)}{h_n'(ka)}h_n(ka); \tag{35}$$

j_n and h_n are the spherical Bessel function of the first kind and the Hankel function of the first kind [29, §10.2], respectively, and a prime indicates the derivative of a function with respect to its argument. Figure 4 shows the magnitude of the modal coefficients as a function of ka. It is apparent from the figure that the spherical array will have poor directivity at the lowest frequencies—such as $ka = 0.2$ which corresponds to 260 Hz for $a = 4.2$ cm—inasmuch as only Y_0 is available for beamforming; amplifying the higher order modes at these frequencies would introduce a great deal of sensor self noise into the beamformer output. From Figure 3 a), however, it is clear that Y_0 is completely isotropic; i.e., it has no directional characteristics and hence provides no improvement in directivity over a single omnidirectional microphone.

The sound field G can be decomposed by the spherical Fourier transform as

$$G_n^m(ka, \Omega_0) = \int_\Omega G(ka, \Omega, \Omega_0) \bar{Y}_n^m(\Omega) d\Omega \tag{36}$$

and the inverse transform is defined as

$$G(ka, \Omega, \Omega_0) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} G_n^m(ka, \Omega_0) Y_n^m(\Omega). \tag{37}$$

The transform (36) can be intuitively interpreted as the decomposition of the sound field into the spherical harmonics illustrated in Figure 3.

Upon substituting the plane wave (34) into (36), we can represent the plane wave in the spherical harmonics domain as

$$G_n^m(ka, \Omega_0) = 4\pi\, i^n\, b_n(ka)\, \bar{Y}_n^m(\Omega_0). \tag{38}$$

In order to understand how beamforming may be performed in the spherical harmonic domain, we need only define the *modal array manifold vector* [30, §5.1.2] as

$$\mathbf{v}(ka, \Omega_0) \triangleq \begin{bmatrix} G_0^0(ka, \Omega_0) \\ G_1^{-1}(ka, \Omega_0) \\ G_1^0(ka, \Omega_0) \\ G_1^1(ka, \Omega_0) \\ G_2^{-2}(ka, \Omega_0) \\ G_2^{-1}(ka, \Omega_0) \\ G_2^0(ka, \Omega_0) \\ \vdots \\ G_N^{-N}(ka, \Omega_0) \\ \vdots \\ G_N^N(ka, \Omega_0) \end{bmatrix}, \tag{39}$$

which fulfills precisely the same role as (1). It is similarly possible to define a noise plus interference vector $\mathbf{N}(ka)$ in spherical harmonic space. Moreover, Yan et al. [31] demonstrated that the covariance matrix for the spherically isotropic noise field in spherical harmonic space can be expressed as

$$\begin{aligned} \mathbf{\Gamma}(ka) \; &= 4\,\pi\sigma_{\text{SI}}^2\, \text{diag}\{|b_0(ka)|^2, -|b_1(ka)|^2, -|b_1(ka)|^2, \\ &\quad -|b_1(ka)|^2, |b_2(ka)|^2, \cdots, (-1)^N |b_N(ka)|^2\}, \end{aligned} \tag{40}$$

where σ_{SI}^2 is the noise power spectral density.

With the changes described above, all of the relations developed in Section II can be applied; the key intuition is that the physical microphones have been replaced by the spherical harmonics which have the very attractive property of the orthonormality as indicated by (33). In particular, the weights of the delay-and-sum beamformer can be calculated as in (3), the MVDR weights as in (6), and the diagonally loaded MVDR weights as in (7); the spherical harmonics super-directive beamformer is obtained by replacing $\mathbf{\Sigma_N}$ in (7) with (40).

A. Discretization

In practice, it is impossible to construct a continuous, pressure sensitive spherical surface; the pressure must be sampled at S discrete points with microphones. The discrete spherical Fourier transform and the inverse transform can be written as

$$G_n^m(ka) = \sum_{s=0}^{S-1} \alpha_s G(ka, \Omega_s) \bar{Y}_n^m(\Omega_s), \tag{41}$$

$$G(ka, \Omega_s) = \sum_{n=0}^{N} \sum_{m=-n}^{n} G_n^m(ka) Y_n^m(\Omega_s), \tag{42}$$

where Ω_s indicates the position of microphone s and α_s is a quadrature constant. Typically N is limited such that $(N+1)^2 \leq S$ to prevent spatial aliasing [9].

Accordingly, the orthonormality condition (33) is approximated by the weighted summation, which causes orthonormality error [28]. In order to alleviate the error caused by discreteness, spatial sampling schemes [10] or beamformer's weights [28] must be carefully designed. In this article, we use a spherical microphone array with 32 equidistantly spaced sensors and set $\alpha_s = 4\pi/S$ in (41) for the experiments described later.

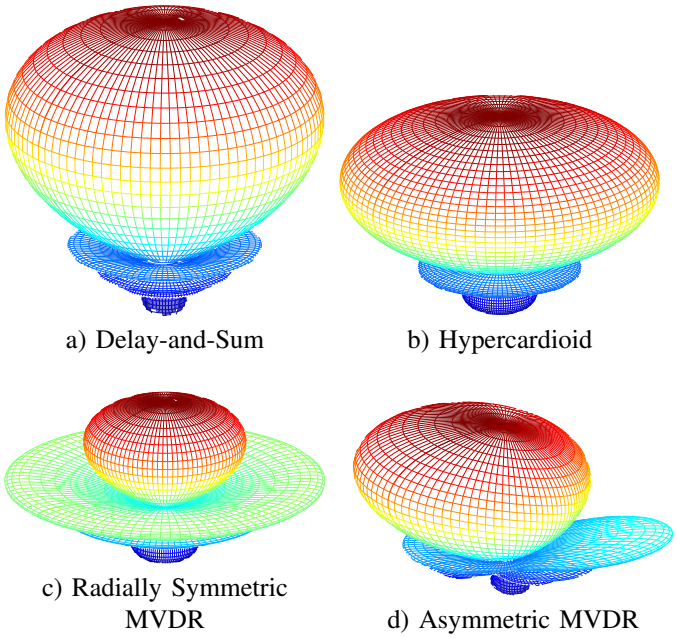
V. Spherical Adaptive Algorithms

In this case we investigate the MVDR beamformer described in Section II-B for spherical arrays. The solution for the MVDR beamforming weights with diagonal loading is given by (7). As discussed in the prior section, in the case of a spherical array, we treat each modal component as a microphone, and apply the beamforming weights directly to the output of each mode. In so doing, we are adhering to the decomposition of the entire beamformer into *eigenbeamformer* followed by a *modal beamformer* as initially proposed by [9], [32].

With formulations of the relevant array manifold vector (39), we can immediately write the solution (3) for the *delay-and-sum* beamformer. Another popular fixed design for spherical array processing is the hypercardioid [9]. In order to illustrate the differences between these several designs, let us now consider the case wherein the look direction is $(\theta, \phi) = (0, 0)$ and there is a single strong interference signal impinging on the array from $(\theta_{\text{I}}, \phi_{\text{I}}) = (\pi/6, 0)$ with a magnitude of $\sigma_{\text{I}}^2 = 10^{-1}$. In this case, the covariance matrix of the array input is

$$\begin{aligned} \Sigma_{\mathbf{X}}(ka) = \mathbf{v}(\Omega, ka)\, \mathbf{v}^H(\Omega, ka) \\ + \sigma_{\text{I}}^2\, \mathbf{v}(\theta_{\text{I}}, \phi_{\text{I}}, ka)\, \mathbf{v}^H(\theta_{\text{I}}, \phi_{\text{I}}, ka). \end{aligned} \tag{43}$$

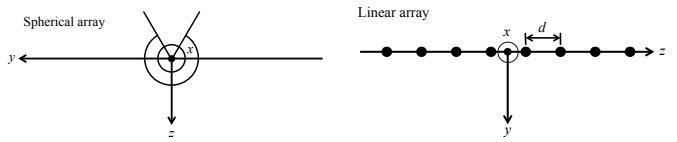
The beampatterns obtained with the delay-and-sum and hypercardioid designs are shown in Figure 5a) and b) respectively. The MVDR design both with and without a radial symmetry constraint are shown in Figure 5c) and d), respectively.



a) Delay-and-Sum b) Hypercardioid

c) Radially Symmetric MVDR d) Asymmetric MVDR

Fig. 5. Spherical beampatterns for $ka = 10.0$: a) Delay-and-sum beampattern; b) Hypercardioid beampattern $H = Y_0 + \sqrt{3}Y_1 + \sqrt{5}Y_2$; c) Symmetric MVDR beampattern obtained with spherical harmonics Y_n for $n = 0, 1, \ldots, 5$, diagonal loading $\sigma_{\mathrm{D}}^2 = 10^{-2}$ for a plane wave interferer $\pi/6$ rad from the look direction; d) Asymmetric MVDR beampattern obtained with spherical harmonics Y_n^m for $n = 0, 1, \ldots, 5$, $m = -n, \ldots, n$, diagonal loading $\sigma_{\mathrm{D}}^2 = 10^{-2}$ for a plane wave interferer $\pi/6$ rad from the look direction.



a) Spherical 32-channel array. b) Conventional, 64-channel, linear array.

Fig. 6. Orientation of the a) Mark IV linear array and b) Eigenmike® spherical array.

VI. Comparative Studies

In this section, we present a set of comparative studies for a conventional linear array and a spherical array. We first compare the arrays on the basis of the theoretical performance metrics introduced in Section III, namely array gain, white noise gain, and directivity index. Thereafter, we compare the arrays on a metric of more direct interest to those researchers on the forefront of distant speech recognition technology, namely, word error rate.

The orientation of the conventional, linear and spherical arrays shown in Figure 6 were used as the basis for evaluating array gain, white noise gain, and directivity index; these configurations were intended to simulate the condition wherein the arrays are mounted at head height for a standing speaker. The acoustic environment we simulated involved a desired speaker, a source of discrete interference—such as a screen projector—somewhat below and to the left of the desired source, and a spherically isotropic noise field—such as might be created by an air conditioning system; the details of the environment, which is equivalent for both arrays, are summarized in Table I. The specific arrays we chose to simulate were the Mark IV linear array and the Eigenmike® spherical array.

Source	Position $\Omega \triangleq (\theta, \phi)$ Mark IV	Eigenmike	Level (dB)
Desired	$(3\pi/8, 0)$	$(\pi/2, -\pi/8)$	0
Discrete Interference	$(3\pi/4, \pi/8)$	$(3\pi/8, \pi/4)$	-10
Diffuse Noise	—	—	-10

TABLE I
Acoustic environment for comparing the Mark IV linear array with the Eigenmike® spherical microphone array.

In Figure 7 are shown the plots of array gain as a function of ka of for the ideal spherical array as well as the discrete arrays with $S = 24$ and 32. From these plots two facts become apparent. Firstly, the MVDR beamformer, as anticipated by the theory presented in Section II-B, provides the highest array gain overall. This was to be expected because minimizing the noise variance is equivalent to maximizing SNR if performed over each individual subband; in order to maximize SNR over the entire subband, the subband signals must be weighted by a Wiener filter prior to their combination. Secondly, the figures for $S = 24$ and 32 indicate that the array gain of the ideal array is reduced when the array must be implemented in hardware with discrete microphones.

Figure 8 shows the white noise gain (WNG) for the ideal spherical array, as well as its discrete counterparts for $S = 24$ and 32. Once more, as predicted by the theory, the uniform (i.e., D&S) beamformer provides the best performance according to this metric. The SD and MVDR beamformers provide substantially lower WNG at low frequencies, but essentially equivalent performance for $ka \geq 30$.

The beampattern is the sensitivity of the array to a plane wave arriving from some direction Ω. By weighting each spherical mode (38) by $\bar{w}_n^m$, the beampattern for the ideal array can be expressed as

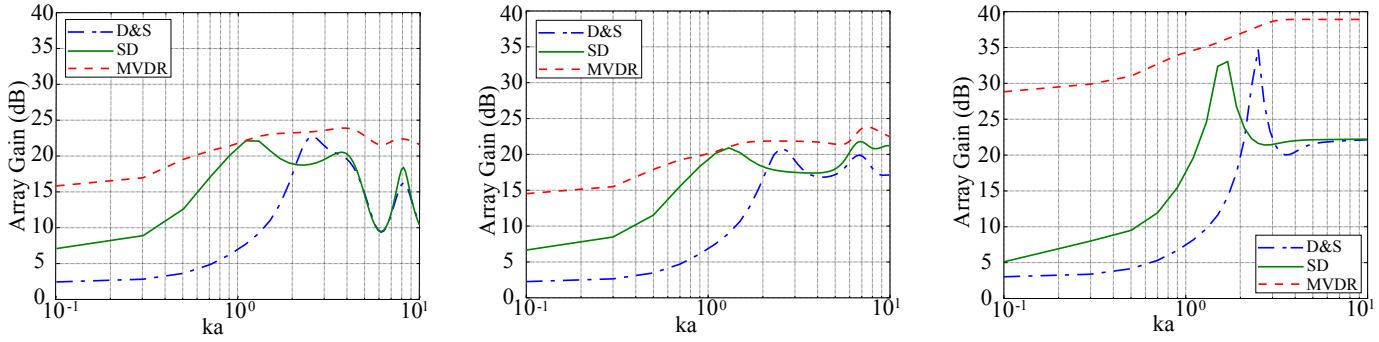
$$B(\Omega, ka) = 4\pi \sum_{n=0}^{N} i^n b_n(ka) \sum_{m=-n}^{n} \bar{w}_n^m \bar{Y}_n^m(\Omega).$$

This implies that the power pattern (27) is given by

$$\begin{aligned} P(\Omega) &\triangleq |B(\Omega, ka)|^2 \qquad (44) \\ &= 16\pi^2 \sum_{n,n'=0}^{\infty} i^n \bar{i}^{n'} b_n(ka) \bar{b}_{n'}(ka) \cdot \\ &\qquad \sum_{m,m'=-n,-n'}^{n,n'} \bar{w}_n^m w_{n'}^{m'} \bar{Y}_n^m(\Omega) Y_{n'}^{m'}(\Omega). \end{aligned}$$

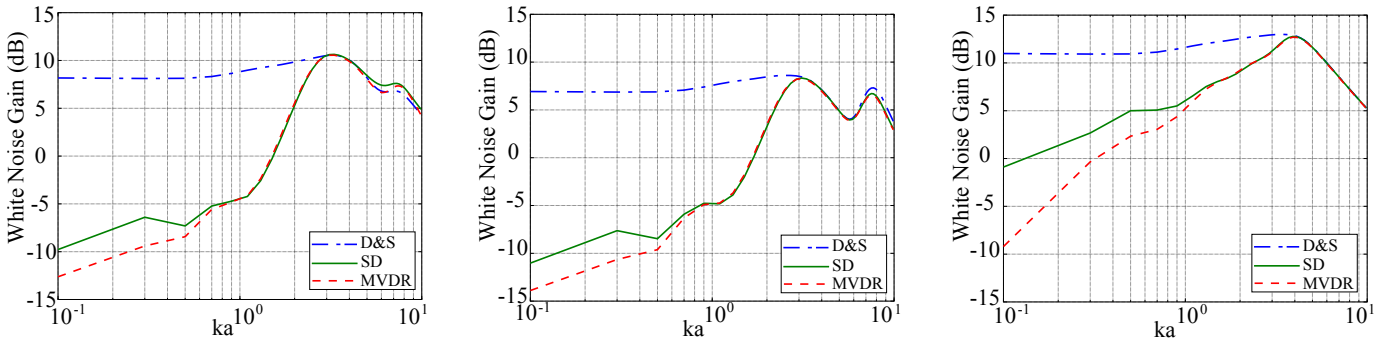
Substituting (44) into (29) and applying (33) then provides [11]

$$\begin{aligned} &\mathrm{DI}_{\mathrm{ideal}}(ka, \mathbf{w}) = \\ &-10 \log_{10} \left\{ 4\pi \sum_{n=0}^{N} |b_n(ka)|^2 \sum_{m=-n}^{m} |w_n^m|^2 \right\}. \qquad (45) \end{aligned}$$



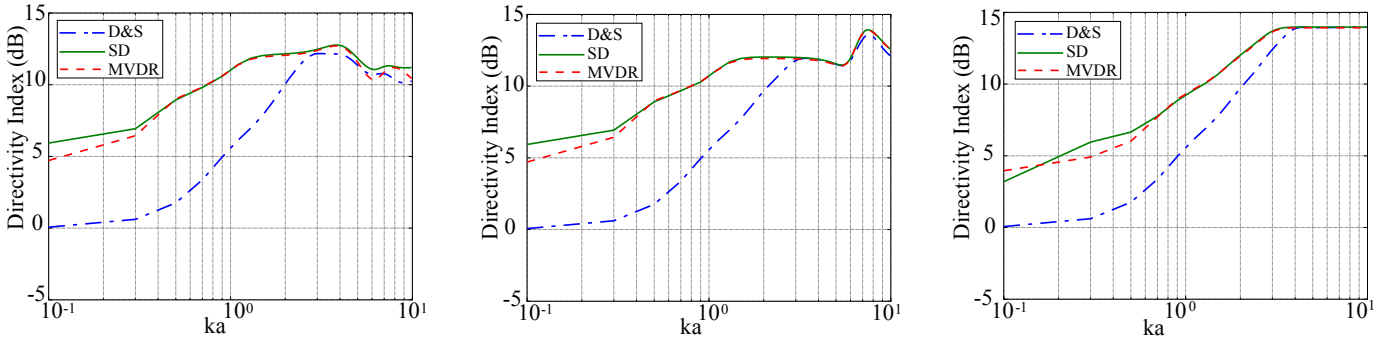
a) $S = 24$ b) $S = 32$ c) Ideal array

Fig. 7. Array gain as a function of ka for a) $S = 24$, b) $S = 32$, and c) the Ideal array.



a) $S = 24$ b) $S = 32$ c) Ideal array

Fig. 8. White noise gain as a function of ka for a) $S = 24$, b) $S = 32$, and c) the Ideal array.



a) $S = 24$ b) $S = 32$ c) Ideal array

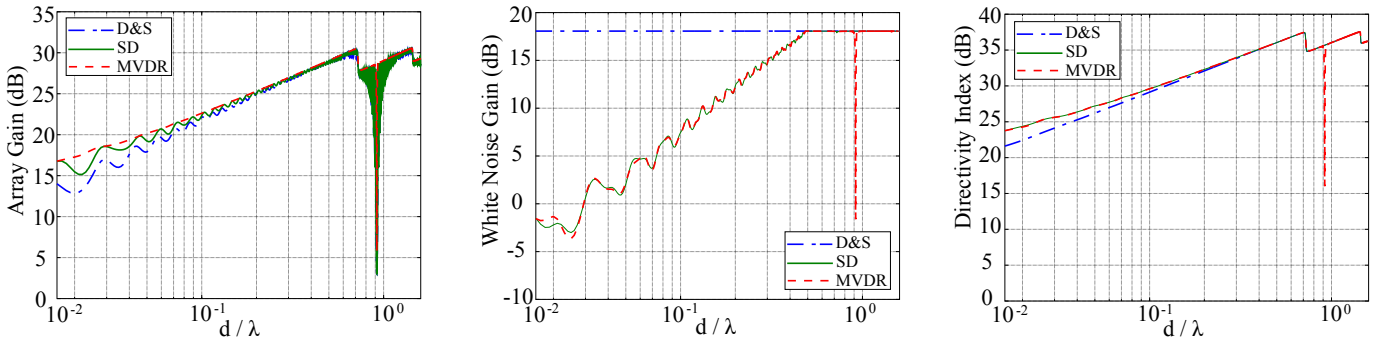
Fig. 9. Directivity index as a function of ka for a) $S = 24$, b) $S = 32$, and c) the Ideal array.

The directivity index as a function of ka for both ideal and discrete arrays is plotted in Figure 9. These figures reveal that—as anticipated by the theory of Section II-C—the superdirective (SD) beamformer provides the highest directivity save in the very low frequency region where the sensor covariance matrix (8) is dominated by the diagonal loading.

Now we come to an equivalent set of plots for the Mark IV linear array; these are shown in Figure 10, where each metric is shown as a function of d/λ, the ratio of intersensor spacing to wavelength. Once more, the MVDR beamformer provides the highest array gain, the D&S beamformer the highest white noise gain, and the superdirective beamformer the highest directivity index. What is unsurprising is that the Mark IV provides a higher array gain than the Eigenmike overall, given its greater number of sensors. What is somewhat surprising is the drastic drop in all metrics just below the point $d/\lambda = 1$; this stems from the fact that this is the point where the first grating lobe crosses the source of discrete interference. A grating lobe cannot be suppressed given that—due to spatial aliasing—it is indistinguishable from the main lobe and hence subject to the distortionless constraint (4).

VII. Comparison of Linear and Spherical Arrays for DSR

As a spherical microphone array has—to the best knowledge of the current authors—never before been applied to DSR,



a) Array gain b) White noise gain c) Directivity index

Fig. 10. a) Array gain, b) white noise gain, and c) directivity index as a function of d/λ for the 64-element, linear Mark IV micophone array with an intersensor spacing of $d = 2$ cm.

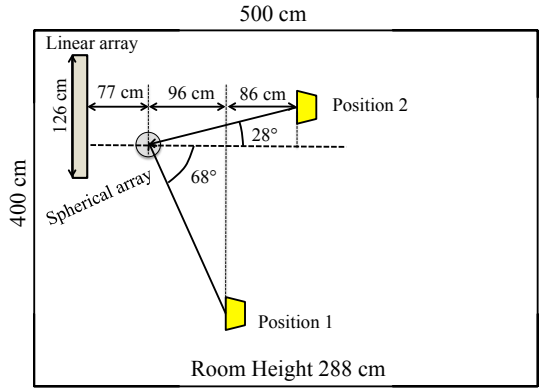


Fig. 11. The layout of the recording room.

Beamforming (BF) Algorithm	Pass (%WER)			
	1	2	3	4
Single array channel (SAC)	47.3	18.9	14.3	13.6
D&S BF with linear array	44.7	17.2	11.1	9.8
SD BF with linear array	45.5	16.4	10.7	9.3
Spherical D&S BF	47.3	16.8	13.0	12.0
Spherical SD BF	42.8	14.5	11.5	10.2
CTM	16.7	7.5	6.4	5.4

TABLE II
WERs FOR EACH BEAMFORMING ALGORITHM IN THE CASE THAT THE ANGLE OF INCIDENCE TO THE ARRAY IS 28^o.

our first step in investigating its suitability for such a task was to capture some prerecorded speech played into a real room through a loudspeaker, then perform beamforming and subsequently speech recognition. Figure 11 shows the configuration of room used for these recordings. As shown in the figure, the loudspeaker was placed in two different positions; the locations of the sensors and loudspeaker were measured with *OptiTrack*, a motion capture system manufactured by *NaturalPoint*. For data capture we used an Eigenmike® which consists of 32 microphones embedded in a rigid sphere with a radius of 4.2 cm; for further details see the website of mh acoustics, http://www.mhacoustics.com. Each sensor of the Eigenmike® is centered on the face of a truncated icosahedron. We simultaneously captured the speech data with a 64-channel, uniform linear Mark IV microphone array with an intersensor spacing of 2 cm for a total aperture length of 126 cm. Speech data from the corpus were used as test material. The test set consisted of 3,241 words uttered by 37 speakers for each recording position. The far-field data was sampled at a rate of 44.1 kHz. The reverberation time T_{60} in the recording room was approximately 525 ms.

We used the speech recognition system and passes described in Kumatani *et al* [12]. Tables II and III show word error rates (WERs) for each beamforming algorithm for the cases wherein the angles of incidence of the target signal to the array were 28^o and 68^o, respectively. As a reference, the WERs obtained with a single array channel (SAC) and the clean data played through the loudspeaker (Clean data) are also reported. It is clear from the tables that every beamforming algorithm provides superior recognition performance to the SAC after the last adapted pass of recognition. It is also clear from the tables that superdirective beamforming with the small spherical array of radius 4.2 cm (Spherical SD BF) can achieve recognition performance very comparable to that obtained with the same beamforming method with the linear array (SD BF with linear array). In the case that the speaker position is nearly in front of the array, superdirective beamforming with the linear array (SD BF with linear array) can still achieve the best result among all the algorithms. This is because of the highest directivity index can be achieved with 64 channels, twice as many as the sensors as in the spherical array. In the other configuration, however, wherein the desired source is at an oblique angle to the array, the spherical superdirective beamformer (Spherical SD BF) provides better results than the linear array because it is able to maintain the same beam pattern regardless of the angle of incidence. In these experiments, spherical D&S beamforming (Spherical D&S BF) could not improve the recognition performance significantly because of its poor directivity.

VIII. Conclusions and Future Directions

This contribution provided an overview of standard beamforming methods for spherical microphone arrays. Addtionally, we compared conventional linear and spherical arrays in

Beamforming (BF) Algorithm	Pass (%WER)			
	1	2	3	4
Single array channel (SAC)	57.8	25.1	19.4	16.6
D&S BF with linear array	53.6	24.3	16.1	13.3
SD BF with linear array	52.6	23.8	16.6	12.8
Spherical D&S BF	57.6	22.7	14.9	13.5
Spherical SD BF	44.8	15.5	11.3	9.7
CTM	16.7	7.5	6.4	5.4

TABLE III
WERS FOR EACH BEAMFORMING ALGORITHM IN THE CASE THAT THE ANGLE OF INCIDENCE TO THE ARRAY IS 68^o.

terms of the theoretical performance metrics typically used in acoustic beamforming. Finally, we presented the results of several distant speech recognition (DSR) experiments comparing linear and spherical arrays under the realistic conditions. The results suggested that the compact spherical microphone array can achieve recognition performance comparable or superior to that of a large linear array. In our view, a key research topic in future will be the efficient integration of various information sources such as video modalities and turn-taking models into a DSR-based dialogue systems.

REFERENCES

[1] M. Omologo, M. Matassoni, and P. Svaizer, "Environmental conditions and acoustic transduction in hands-free speech recognition," *Speech Communication*, vol. 25, pp. 75–95, 1998.

[2] Matthias Wölfel and John McDonough, *Distant Speech Recognition*, Wiley, New York, 2009.

[3] Maurizio Omologo, "A prototype of distant-talking interface for control of interactive TV," in *Proc. Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, Pacific Grove, CA, 2010.

[4] M. Lincoln, I. McCowan, I. Vepa, and H. K. Maganti, "The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments," in *Proc. IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2005, pp. 357–362.

[5] John McDonough, Kenichi Kumatani, Tobias Gehrig, Emilian Stoimenov, Uwe Mayer, Stefan Schacht, Matthias Wölfel, and Dietrich Klakow, "To separate speech!: A system for recognizing simultaneous speech," in *Proc. MLMI*, 2007.

[6] John McDonough and Matthias Wölfel, "Distant speech recognition: Bridging the gaps," in *Proc. IEEE Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Trento, Italy, 2008.

[7] Michael Seltzer, "Bridging the gap: Towards a unified framework for hands-free speech recognition using microphone arrays," in *Proc. HSCMA*, Trento, Italy, 2008.

[8] H. L. Van Trees, *Optimum Array Processing*, Wiley-Interscience, New York, 2002.

[9] Jens Meyer and Gary W. Elko, "Spherical microphone arrays for 3D sound recording," in *Audio Signal Processing for Next–Generation Multimedia Communication Systems*, pp. 67–90. Kluwer Academic, Boston, MA, 2004.

[10] Boaz Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 135–143, 2005.

[11] Shefeng Yan, Haohai Sun, U. Peter Svensson, Xiaochuan Ma, and Jens M. Hovem, "Optimal modal beamforming for spherical microphone arrays," *IEEE Trans. on Audio Speech Language Processing*, vol. 19, no. 2, pp. 361–371, February 2011.

[12] Kenichi Kumatani, Takayuki Arakawa, Kazumasa Yamamoto, John McDonough, Bhiksha Raj, Rita Singh, and Ivan Tashev, "Microphone array processing for distant speech recognition: Towards real-world deployment," in *Proc. APSIPA ASC*, Hollywood, CA, December 2012.

[13] Kenichi Kumatani, John McDonough, and Bhiksha Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Processing Magazine*, September 2012.

[14] John McDonough and Kenichi Kumatani, "Microphone arrays," in *Techniques for Noise Robustness in Automatic Speech Recognition*, Tuomas Virtanen, Rita Singh, and Bhiksha Raj, Eds., chapter 6. Wiley, London, November 2012.

[15] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Audio, Speech and Language Processing*, vol. ASSP-35, pp. 1365–1376, 1987.

[16] M. Brandstein and D. Ward, Eds., *Microphone Arrays*, Springer Verlag, Heidelberg, Germany, 2001.

[17] Simon Haykin, *Adaptive Filter Theory*, Prentice Hall, New York, fourth edition, 2002.

[18] Iain A. McCowan and Hervé Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech Audio Processin*, vol. 11, pp. 709–716, 2003.

[19] Rita Singh, Kenichi Kumatani, John McDonough, and Chen Liu, "Signal-separation-based array postfilter for distant speech recognition," in *Proc. Interspeech*, Portland, OR, 2012.

[20] Kenichi Kumatani, Bhiksha Raj, Rita Singh, and John McDonough, "Microphone array post-filter based on spatially-correlated noise measurements for distant speech recognition," in *Proc. Interspeech*, Portland, OR, 2012.

[21] Tobias Wolff and Markus Buck, "A generalized view on microphone array postfilters," in *Proc. International Workshop on Acoustic Signal Enhancement*, Tel Aviv, Israel, 2010.

[22] Claude Marro, Yannick Mahieux, and K. Uwe Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. Speech Audio Process.*, vol. 6, pp. 240–259, 1998.

[23] Kenichi Kumatani, John McDonough, Dietrich Klakow, Philip N. Garner, and Weifeng Li, "Adaptive beamforming with a maximum negentropy criterion," *IEEE Trans. Audio, Speech, and Language Processing*, August 2008.

[24] John McDonough, Matthias Wölfel, and Alex Waibel, "On maximum mutual information speaker–adapted training," *Computer Speech and Language*, December 2007.

[25] E. N. Gilbert and S. P. Morgan, "Optimum design of antenna arrays subject to random variations," *Bell Syst. Tech. J.*, vol. 34, pp. 637–663, May 1955.

[26] Joerg Bitzer and K. Uwe Simmer, "Superdirective microphone arrays," in *Microphone Arrays*, M. Branstein and D. Ward, Eds., pp. 19–38. Springer, Heidelberg, 2001.

[27] Earl G. Williams, *Fourier Acoustics*, Academic Press, San Diego, CA, USA, 1999.

[28] Zhiyun Li and Ramani Duraiswami, "Flexible and optimal design of spherical microphone arrays for beamforming," *IEEE Trans. Speech Audio Process.*, vol. 15, pp. 2007, 702–714.

[29] Frank W. J. Olver and L. C. Maximon, "Bessel functions," in *NIST Handbook of Mathematical Functions*, Frank W. J. Olver, Daniel W. Lozier, Ronald F. Boisvert, and Charles W. Clark, Eds. Cambridge University Press, New York, NY, 2010.

[30] Heinz Teutsch, *Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition*, Springer, Heidelberg, 2007.

[31] Shefeng Yan, Haohai Sun, U. Peter Svensson, Xiaochuan Ma, and J. M. Hovem, "Optimal modal beamforming for spherical microphone arrays," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 2, pp. 361–371, 2011.

[32] Jens Meyer and Gary W. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proc. ICASSP*, Orlando, FL, May 2002.