1

Microphone Arrays

John McDonough
Carnegie Mellon University

Kenichi Kumatani
Disney Research, Pittsburgh

This contribution takes as its objective the class of techniques suitable for performing speech recognition, not on the signal capture by a single microphone, but on that obtained by combining the signals from several microphones. The techniques discussed here differ from those presented in Chapter **??** in that they are based on the pair of assumptions that:

1. The geometry of the array of microphones is fixed and known;
2. The position of the active speakers relative to the array are known or can be accurately estimated.

Such techniques—known collectively as *beamforming*—have been the subject of intense interest in recent years within the acoustic array processing research community. Unfortunately, such techniques have been largely ignored in the mainstream automatic speech recognition field, although this may rapidly change given the recent release and widespread popularity of the Microsoft Kinect® platform. The simplest of beamforming algorithms, the *delay-and-sum beamformer*, uses only this geometric knowledge—i.e., the arrangement of the microphones and the speaker's position—to compensate for the time delays of the signals arriving at each sensor and then additively combine them. More sophisticated *adaptive beamformers* minimize the total output power of the array under the constraint that the desired source must be unattenuated. The conventional adaptive beamforming algorithms attempt to minimize a quadratic optimization criterion related to signal-to-noise ratio. Recent research has revealed, however, that such quadratic criteria are not optimal for acoustic beamforming of human speech. Hence, we also present beamformers based on non-conventional optimization criteria, that have appeared more recently in the literature. In particular, recent research has revealed that useful optimization criteria can be devised by attempting to restore the non-Gaussian statistical characteristics present in uncorrupted or "clean" speech. As these characteristics are diminished through the introduction of noise or reverberation, the use of adaptive beamforming techniques to restore the original statistical characteristics reduces the effect of these distortions, and hence improves speech recognition performance.

A second research trend upon which we will report is the growing use of *spherical* microphone arrays. The literature on array processing with spherical arrays differs from the "conventional" array processing literature in that it attempts to explicitly account for diverse acoustic phenomena, namely, the diffraction of sound around a solid sphere, as well its scattering from such an object. While diffraction and scattering are present in all acoustic array processing applications, the conventional literature takes them into account only through the calculation of second order statistics between pairs of sensors. In the spherical array literature, on the other hand, these effects are incorporated into the theoretical analysis.

A third trend upon which we report is the combination of adaptive beamforming techniques developed for conventional arrays with the acoustic theory developed for spherical arrays. This all important research direction, which has only very recently appeared, will, in the opinion of the current authors, dominate the field in the coming years and decades.

The balance of this contribution is organized as follows. We begin in Section 1.1 by discussing speaker tracking based on the use of Bayesian filters. In Section 1.2 we review the basics of the conventional array processing literature. Beginning with the theory of linear apertures, we investigate the effects of processing with discrete arrays as well as array steering. Two important concepts introduced in this section are those of poor low-frequency directivity, which arises from the finite extent of an array, and high-frequency spatial aliasing, which arises from the necessity of sampling an aperture at discrete points. Our discussion of adaptive array processing begins in Section 1.3. This includes both array processing based on second order statistics, as introduced in Section 1.3.1, as well as that based on higher order statistics or *non-Gaussian* criteria, as discussed in Section 1.3.8. Other topics covered in Section 1.3 include theoretical models for noise fields in Section 1.3.2, subband analysis and synthesis for adaptive filtering and beamforming in Section 1.3.3, beamforming performance criteria in Section 1.3.4, the generalized sidelobe canceller in Section 1.3.5 as well as its recursive implementation in Section 1.3.6, and other conventional beamformers in Section 1.3.7. The first set of distant speech recognition results is presented in Section 1.3.10; these compare the performance of several different beamforming optimization criteria. Section 1.4 takes up our presentation of spherical array processing; we discuss acoustic diffraction and scattering, as well as their effects on the sensitivity of a spherical array to plane waves. We also introduce the concept of decomposing a sound field into spherical harmonics, which can be used for beamforming much like the output of a single microphone is used in conventional beamforming techniques. Section 1.5 describes how beamforming techniques based on the second order statistics discussed in Sections 1.3.1 can be profitably applied to spherical array processing. A comparison of conventional, linear and spherical arrays is presented in Section 1.6 based on the beamforming performance criteria defined in Section 1.3.4. Thereafter our second set of distant speech recognition results comparing the two arrays is presented. Finally, in Section 1.8 we present our conclusions as well as our suggestions for further reading.

1.1 Speaker Tracking

Before beamforming can be effectively used to enhance the speech of a desired speaker in a distant speech recognition application, the speaker's position, denoted as $\mathbf{x}$, relative to the microphone array must be known or estimated. Hence, let us begin our discussion by briefly

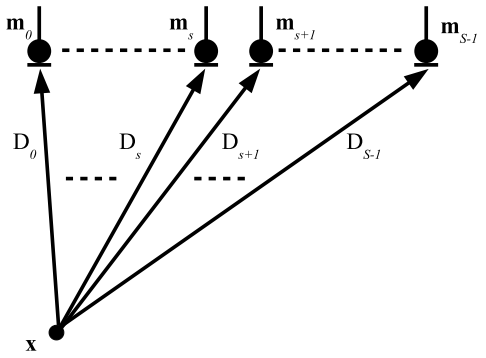


Figure 1.1 Positions of the microphones $\{\mathbf{m}_s\}$ and speaker $\mathbf{x}$, as well as the distances between them $\{D_s\}$.

examining how such an estimation can be performed.

The *time delay of arrival* (TDOA) between the microphones at positions $\mathbf{m}_1$ and $\mathbf{m}_2$ can be expressed as

$$T(\mathbf{m}_1, \mathbf{m}_2, \mathbf{x}) \triangleq \frac{\|\mathbf{x} - \mathbf{m}_1\| - \|\mathbf{x} - \mathbf{m}_2\|}{c} \tag{1.1}$$

where c is the speed of sound, which is approximately 344 m/s at sea level. The definition (1.1) can be rewritten as

$$T(\mathbf{m}_m, \mathbf{m}_n, \mathbf{x}) \triangleq \frac{D_m - D_n}{c}, \tag{1.2}$$

where

$$D_n \triangleq \|\mathbf{x} - \mathbf{m}_n\| \ \ \forall \ \ n = 0, \ldots, S-1 \tag{1.3}$$

is the distance from the speaker to the microphone located at $\mathbf{m}_n$ and S is the total number of microphones, as shown in Figure 1.1.

Let $\hat{\tau}_{mn}$ denote the *observed* TDOA for the mth and nth microphones. The TDOA can be observed or estimated with a variety of well-known techniques. Perhaps the most popular method involves the *phase transform* (PHAT) (Carter 1981), a variant of the *generalized cross-correlation* (GCC), which can be expressed as

$$\rho_{mn}(\tau) \triangleq \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{Y_m(e^{j\omega\tau}) Y_n^*(e^{j\omega\tau})}{|Y_m(e^{j\omega\tau}) Y_n^*(e^{j\omega\tau})|} \, e^{j\omega\tau} \, d\omega, \tag{1.4}$$

where $Y_n(e^{j\omega\tau})$ denotes the short-time Fourier transform of the signal arriving at the nth sensor in the array (Omologo and Svaizer 1994). The definition of the GCC in (1.4) follows directly from the frequency domain calculation of the cross-correlation of two sequences. The normalization term $\left|Y_m(e^{j\omega\tau}) Y_n^*(e^{j\omega\tau})\right|$ in the denominator of the integrand is intended to weight all frequencies equally. It has been shown that such a weighting leads to more robust TDOA estimates in noisy and reverberant environments (DiBiase et al. 2001). Once $\rho_{mn}(\tau)$ has been calculated, the TDOA estimate is obtained from

$$\hat{\tau}_{mn} = \max_{\tau} \ \rho_{mn}(\tau). \tag{1.5}$$

In other words, the "true" TDOA is taken as that which maximizes the value of the PHAT. As (1.5) is typically calculated with an inverse *discrete Fourier transform* (DFT), a parabolic interpolation is often performed to overcome the granularity in the estimate due to the digital sampling interval (Omologo and Svaizer 1994). Usually $Y_n(e^{j\omega_k})$ appearing in (1.4) is calculated with a Hamming analysis window of 15 to 25 ms in duration (DiBiase et al. 2001).

Let us assume that the microphones are divided into a number S_2 of distinct microphone pairs. Consider two microphones located at $\mathbf{m}_{s1}$ and $\mathbf{m}_{s2}$ comprising the sth microphone pair, and once more define the TDOA as in (1.2), where $\mathbf{x}$ represents the position of an active speaker, and define $T_s(\mathbf{x}) \triangleq T(\mathbf{m}_{s1}, \mathbf{m}_{s2}, \mathbf{x})$. Source localization based on the maximum likelihood criterion (Kay 1993) proceeds by minimizing the error function

$$\epsilon(\mathbf{x}) = \sum_{s=0}^{S_2-1} \frac{\left[\hat{\tau}_s - T_s(\mathbf{x})\right]^2}{\sigma_s^2}, \tag{1.6}$$

where σ_s^2 denotes the error covariance associated with this observation, and $\hat{\tau}_s$ is the observed TDOA as in (1.4) and (1.5).

Although (1.6) implies we should find that $\mathbf{x}$ minimizing the instantaneous error criterion, we would be better advised to attempt to minimize such an error criterion over a series of time instants. In so doing, we exploit the fact that the speaker's position cannot change instantaneously; thus, both the present and past TDOA estimates are potentially useful in estimating a speaker's current position. Klee et al. (2005) proposed to recursively minimize the least square error position estimation criterion (1.6) with a variant of the *extended Kalman filter* (EKF). This was achieved by first associating the *state* $\mathbf{x}_k$ of the EKF with the speaker's position at time k, and the kth observation with a vector of TDOAs. In keeping with the formalism of the EKF, Klee et al. then postulated a *state* and *observation equation*,

$$\mathbf{x}_k = \mathbf{F}_{k|k-1}\mathbf{x}_{k-1} + \mathbf{u}_{k-1}, \text{ and} \tag{1.7}$$

$$\mathbf{y}_k = \mathbf{H}_{k|k-1}(\mathbf{x}_k) + \mathbf{v}_k, \tag{1.8}$$

respectively, where

- $\mathbf{F}_{k|k-1}$ denotes the *transition matrix*,
- $\mathbf{u}_{k-1}$ denotes the *process noise*,
- $\mathbf{H}_{k|k-1}(\mathbf{x})$ denotes the vector-valued *observation function*, and
- $\mathbf{v}_k$ denotes the *observation noise*.

The unobservable state $\mathbf{x}_k$ is to be inferred from the sequence $\mathbf{y}_k$ of observations. The process $\mathbf{u}_k$ and observation $\mathbf{v}_k$ noises are unknown, but both have zero-mean Gaussian pdfs and known covariance matrices. Associating $\mathbf{H}_{k|k-1}(\mathbf{x})$ with the TDOA function (1.1) with one component per microphone pair, it is straightforward to calculate the appropriate linearization about the current state estimate required by the EKF (Wölfel and McDonough 2009, §10.2),

$$\bar{\mathbf{H}}_k(\mathbf{x}) \triangleq \nabla_{\mathbf{x}}\mathbf{H}_{k|k-1}(\mathbf{x}). \tag{1.9}$$

By assumption $\mathbf{F}_{k|k-1}$ is known, so that the *predicted state estimate* is obtained from

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{F}_{k|k-1}\hat{\mathbf{x}}_{k-1|k-1}, \tag{1.10}$$

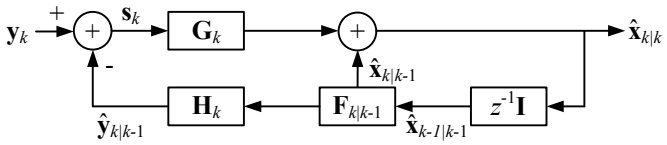


Figure 1.2 Predictor-corrector structure of the Kalman filter.

where $\hat{\mathbf{x}}_{k-1|k-1}$ is the *filtered state estimate* from the prior time step; the calculation of $\hat{\mathbf{x}}_{k|k-1}$ as in (1.10) is known as *prediction*. Let us define the *innovation* as

$$\mathbf{s}_k \triangleq \mathbf{y}_k - \mathbf{H}_{k|k-1}\left(\hat{\mathbf{x}}_{k|k-1}\right).$$

The innovation is called as such because it represents the component of the response of the system that could not be predicted from the state equation (1.7). The new filtered state estimate is calculated from

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{G}_k\,\mathbf{s}_k, \tag{1.11}$$

where $\mathbf{G}_k$ denotes the *Kalman gain*, which can be calculated through a well-known recursion (Wölfel and McDonough 2009, §4.3). A block diagram illustrating the prediction and correction steps in the state estimate update of a conventional Kalman filter is shown in Figure 1.2.

1.2 Conventional Microphone Arrays

We will now analyze the characteristics of conventional apertures and arrays. As we will learn in Section 1.4, several of these characteristics are shared by the less conventional spherical apertures and arrays.

The relationship between the spherical coordinates (r, θ, ϕ) and Cartesian coordinates (x, y, z) is shown in Figure 1.3; the *polar angle* θ and *azimuth* ϕ are measured from the z- and x-axes, respectively, and have ranges $0 \leq \theta \leq \pi$ and $-\pi \leq \phi \leq \pi$ where $\phi = \pi/2$ corresponds to the y-axis. In the figure, a *plane wave* is impinging on an array of microphones located along the x-axis; the vector $\mathbf{a}$ indicates the *direction of arrival* of the plane wave. A plane wave is named as such because any locus of constant phase—or *wavefront*—is a plane; the plane wave assumption is most accurate when the sources are relatively distant from the array as compared to the *aperture length*, whicy by definition is the maximum physical extent of the aperture.

Before taking up the case of conventional microphone arrays, let us consider the *linear aperture* of length L shown in Figure 1.4. The unit normal vector perpendicular to the wavefront can be expressed in Cartesian coordinates as

$$\mathbf{a} = -\begin{bmatrix} \sin\theta\cos\phi \\ \sin\theta\sin\phi \\ \cos\theta \end{bmatrix}. \tag{1.12}$$

The *vector wavenumber*,

$$\mathbf{k} \triangleq \frac{2\pi}{\lambda}\mathbf{a}, \tag{1.13}$$

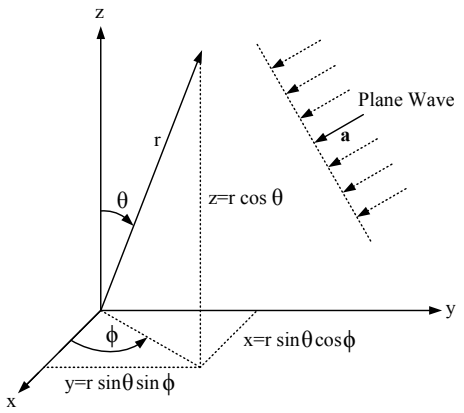


Figure 1.3 Relation between the spherical coordinates (r, θ, ϕ) and Cartesian coordinates (x, y, z).

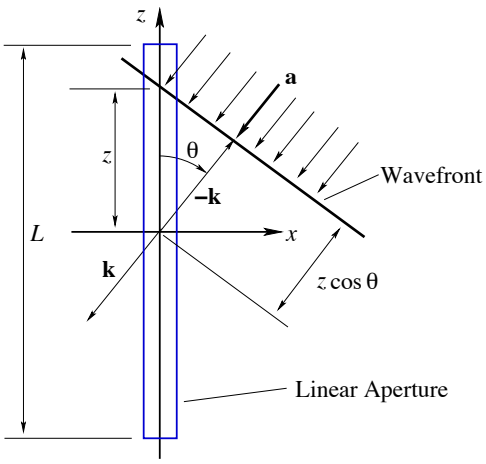


Figure 1.4 A plane wave with normal vector $\mathbf{a}$ and wavenumber $\mathbf{k}$ impinging on a linear aperture of length L lying along the z-axis.

where λ is the length of the propagating wave, indicates both the direction of arrival and frequency of the propagating wave; the direction of arrival is given by $\mathbf{a} \triangleq \mathbf{k}/|\mathbf{k}|$, while the *scalar wavenumber*—defined as

$$k \triangleq \|\mathbf{k}\| = \frac{2\pi}{\lambda} = \frac{\omega}{c}, \tag{1.14}$$

where c is the speed of sound—is the angular frequency of the plane wave. Both k and $\mathbf{k}$ are often referred to as simply the *wavenumber*; we will also adopt this practice where the difference between the two is clear from context. For an arbitrary point $\mathbf{x}$, the TDOA with respect to the origin of the coordinate system is

$$\tau(\mathbf{x}) = \frac{\mathbf{a}^T\mathbf{x}}{c}. \tag{1.15}$$

Assuming now that all points on the linear aperture lie on the z-axis, then (1.12) and (1.15) imply

$$\tau(z) = -\frac{z\cos\theta}{c} = -\frac{uz}{c}, \tag{1.16}$$

where $u \triangleq \cos\theta$ is the *direction cosine* for the z-axis. The component of $\mathbf{k}$ along the z-axis is given by

$$k_z \triangleq -\|\mathbf{k}\|\cos\theta = -\frac{\omega}{c}u = -\frac{2\pi}{\lambda}u. \tag{1.17}$$

From (1.16) and (1.17) it then follows

$$\omega\tau(z) = k_z z. \tag{1.18}$$

Consider now a narrow band source signal $f(t)$ with spectrum $F(\omega)$. Given that a delay $\tau(z)$ in the time domain corresponds to a linear phase shift $e^{-i\tau(z)\omega}$ in the frequency domain, the Fourier transform of the signal component arriving at point z can be expressed as

$$F(\omega, k_z, z) = F(\omega)e^{-i\tau(z)\omega} = F(\omega)e^{-ik_z z}, \tag{1.19}$$

where [1] $i \triangleq \sqrt{-1}$. If the signal components arriving along the entire aperture are *weighted* with a function $w_{\text{a}}^*(z)$ and then *combined*, then the result is the *frequency wavenumber response function*,

$$\Upsilon(\omega, k_z) \triangleq \int_{-\infty}^{\infty} w_{\text{a}}^*(z)e^{-ik_z z}\, dz. \tag{1.20}$$

Let us initially assume that

$$w_{\text{a}}(z) = \frac{1}{L}\begin{cases}1, & \forall \ -L/2 \le z \le L/2,\\ 0, & \text{otherwise.}\end{cases} \tag{1.21}$$

Substituting (1.21) into (1.20), we find

$$\Upsilon(\omega, k_z) = \int_{-L/2}^{L/2} e^{-ik_z z}\, dz = \text{sinc}\left(\frac{L}{2}k_z\right),$$

where

$$\text{sinc}(x) \triangleq \frac{\sin x}{x}. \tag{1.22}$$

Equivalently, given that $\text{sinc}(x)$ is an even function,

$$\Upsilon(\omega, k_z) = \text{sinc}\left(-\frac{L}{2}\cdot\frac{2\pi}{\lambda}u\right) = \text{sinc}\left(\frac{\pi L}{\lambda}\cdot u\right). \tag{1.23}$$

A plot of $\Upsilon(\omega, k_z)$ for several values of L/λ is shown in Figure 1.5. In order to properly analyze the curves in the figure, we must introduce several new terms. With our initial analysis, we strive primarily for intuition rather than mathematical precision; the latter will follow in subsequent sections. First of all, as the x-axis of the plot in Figure 1.6 corresponds to $u = \cos\theta$, we recognize that these curves represent the sensitivity of the array to plane waves impinging from various directions. In general we will refer to such curves as *beampatterns*. The maximum sensitivity for all beampatterns is achieved at $u = 0$, which corresponds to $\theta = \pi/2$. We will refer to this angle of maximum sensitivity as the *look direction*, because it will presumably align with the direction of the desired source. All beampatterns attain a value of unity in the look direction, which implies that a plane wave impinging from this direction is neither amplified nor attenuated; we will refer to this condition by saying that any beampattern fulfilling it satisfies the *distortionless constraint* in the look direction. We will refer to the broad lobe around the look direction as the *main lobe*, and the smaller lobes on either side of the main lobe as *sidelobes*. Finally, we will refer to the capacity of a given beampattern to maximize the ratio of its sensitivity in the look direction to its average sensitivity over all directions as its *directivity*; high directivity is associated with

[1] For present purposes, we break with the signal processing convention of defining $j \triangleq \sqrt{-1}$, as j must be reserved to denote the spherical Bessel function in the sequel.

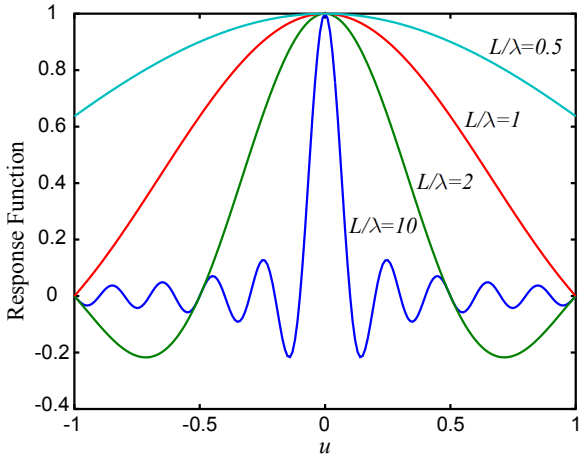


Figure 1.5 Beampatterns for the linear aperture with $L/\lambda = 0.5, 1, 2, 10$.

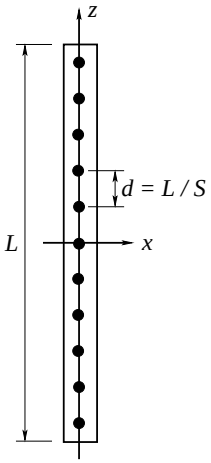


Figure 1.6 A linear aperture of length L and its approximation with an array of $S = 11$ elements with a uniform spacing of $d = L/S$.

focussing on a desired signal impinging from the look direction while suppressing noise and interference from other directions.

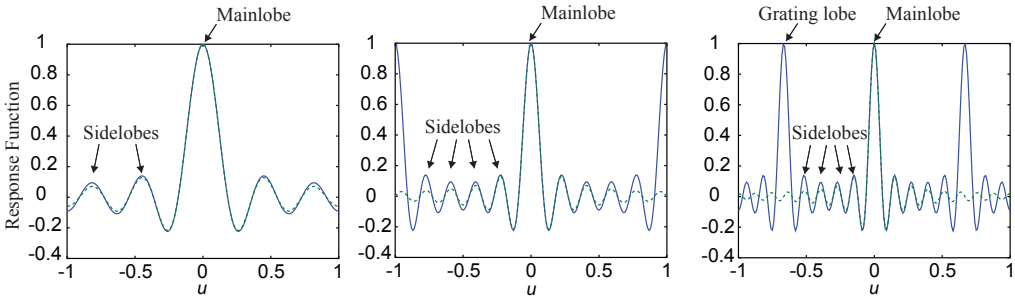
Now that we have equipped ourselves with the proper vocabulary, we can proceed with the analysis of the beampatterns in Figure 1.5. The figure indicates that for very low frequencies in which $L \leq \lambda$, the directivity of the linear aperture is poor. The directivity, however, improves with increasing frequency. Clearly, all beampatterns satisfy the distortionless constraint for a look direction of $u = 0$. The size of the main lobe grows broader with decreasing frequency and increasing wavelength. For higher frequencies with shorter wavelengths, the beampattern exhibits a marked sibelobe structure. As we will learn in Section 1.3 more advanced adaptive beamforming algorithms attempt to reduce the effects of ambient noise and interfering signals by controlling the structure of these sidelobes, a process known as *null steering*.

As a uniformly sensitive aperture is difficult or impossible to construct, let us consider sampling the aperture at S points

$$z_s = \left(s - \frac{S-1}{2} \right) d \ \forall \ s = 0, 1, \ldots, S-1, \tag{1.24}$$

where $d \triangleq L/S$ is the *intersensor spacing* of the array elements as shown in Figure 1.6. This sampling is accomplished by defining the *sampled* sensitivity function

$$w_s(z) \triangleq \frac{1}{S} \sum_{s=0}^{S-1} \delta(z - z_s). \tag{1.25}$$



a) $d/\lambda = \frac{1}{2}$ b) $d/\lambda = 1$ c) $d/\lambda = \frac{3}{2}$

Figure 1.7 Beampatterns for the linear aperture (dotted line) and linear array (solid line) with $S = 11$ and a) $d/\lambda = \frac{1}{2}$, b) $d/\lambda = 1$, and c) $d/\lambda = \frac{3}{2}$.

Substituting (1.25) into (1.20), provides

$$\Upsilon_{\mathrm{s}}(\omega, k_z) = \frac{1}{S} \exp\left\{ik_z d\left(\frac{S-1}{2}\right)\right\} \sum_{s=0}^{S-1} e^{-ik_z sd},$$

which can be readily simplified to (Wölfel and McDonough 2009, §13.1.3)

$$\Upsilon_{\mathrm{S}}(\omega, k_z) = \frac{1}{S} \cdot \frac{\sin\left(S\frac{d}{2}k_z\right)}{\sin\left(\frac{d}{2}k_z\right)} = \mathrm{sinc}_S\left(\frac{d}{2} \cdot \frac{2\pi}{\lambda} \cdot u\right) = \mathrm{sinc}_S\left(\frac{\pi d}{\lambda} \cdot u\right), \tag{1.26}$$

where

$$\mathrm{sinc}_S(x) \triangleq \frac{1}{S} \cdot \frac{\sin Sx}{\sin x}.$$

Note that unlike (1.23), the beampattern (1.26) is *periodic* with period λ/d. Beampatterns $\Upsilon_{\mathrm{S}}(\omega, k_z)$ for several values of d/λ are shown in Figure 1.7. From the figure, it is apparent that for $d/\lambda \leq 1/2$, the behavior of the array is a very good approximation of that of the continuous aperture throughout the entire working range $-1 \leq u \leq 1$. On the other hand, while the behavior of the main lobe around $u =$ is good for $d/\lambda = 1, 3/2$, large spurious lobes with the same magnitude as the main lobe arise at points well-removed from the look direction; these are known as *grating lobes*.

Clearly the look direction for the beampatterns in Figures 1.5 and 1.7 is given by $(\theta_{\mathrm{L}}, \phi_{\mathrm{L}}) = (\pi/2, 0)$, which is typically referred to as *broadside*. Setting the look direction to broadside is achieved with a uniform weighting of the linear aperture as in (1.21), or the uniform weighting of the sensor outputs in (1.25). The process of setting the look direction is known as *beam steering* or simply *steering*. The look direction can readily be set to any desired direction $k = k_{\mathrm{L}}$ by setting the sensor weights to

$$w_{\mathrm{s}}(z; k_{\mathrm{L}}) \triangleq \frac{1}{S} \sum_{s=0}^{S-1} e^{-ik_{\mathrm{L}}d} \delta(z - z_s). \tag{1.27}$$

Doing so yields the beampattern

$$B(k_z, \omega; k_\mathrm{L}) \triangleq \mathbf{v_k}^H(k_\mathrm{L})\mathbf{v_k}(k_z), \tag{1.28}$$

where the *array manifold vector* is defined as

$$\mathbf{v_k}(k_z) \triangleq \left[e^{i\left(\frac{S-1}{2}\right)k_z d} \quad e^{i\left(\frac{S-1}{2}-1\right)k_z d} \quad \dots \quad e^{-i\left(\frac{S-1}{2}\right)k_z d}\right]^T. \tag{1.29}$$

The array manifold vector is nothing more than a vector of phase shifts induced by the propagation delay for each sensor.

While the *visible region* is by definition $-1 \leq u \leq 1$, it is customary to conceptualize u as extending over the entire real line. This is done to facilitate the visualization of grating lobes moving into the visible region as the result of beam steering. The effect of steering is shown in Figure 1.8, from which it is apparent that the grating lobes recur at regular intervals whether or not they are within the visible region. From the figure it is apparent that steering

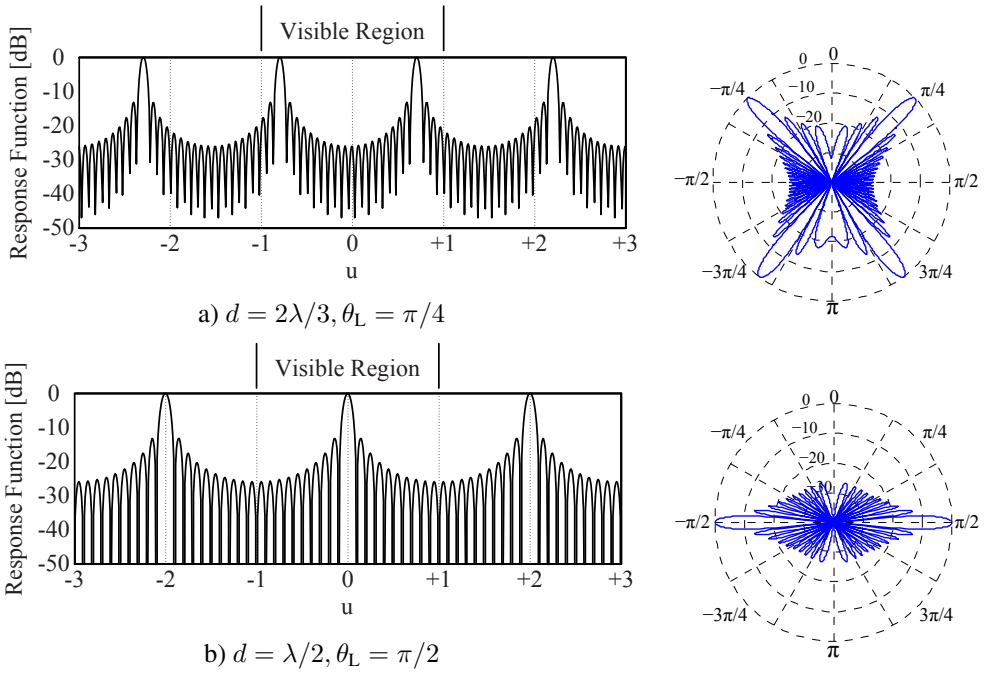


Figure 1.8 Effect of steering on the grating lobes for $S = 20$ plotted in Cartesian and polar coordinates.

can cause grating lobes to enter the beampattern for $d > \lambda/2$. This phenomenon is known as *spatial aliasing*, and occurs when the propagating wave is not sampled sufficiently often *in space*. The *half wavelength rule* (Wölfel and McDonough 2009, §13.1.4) states that avoiding spatial aliasing—even when steering over the entire front half plane—requires

$$\frac{d}{\lambda} \leq \frac{1}{2},$$

which is to say, the wave must be sampled as least twice along its length. This rule is analogous to the *Nyquist sampling theorem* (Oppenheim and Schafer 2010, §4) from conventional signal processing.

The beampattern (1.28) leads to the definition of the *delay-and-sum* beamformer weights as

$$\mathbf{w}_{\mathrm{DS}}^{H}(\omega, k_z) = \frac{\mathbf{v}_{\mathbf{k}}^{H}(\omega, k_z)}{\mathbf{v}_{\mathbf{k}}^{H}(\omega, k_z)\mathbf{v}_{\mathbf{k}}(\omega, k_z)}. \tag{1.30}$$

The delay-and-sum beamformer is the simplest fixed design that satisfies the *distortionless constraint*, as

$$\mathbf{w}^{H}\mathbf{v}_{\mathbf{k}} = 1, \tag{1.31}$$

where for convenience the dependence on (ω, k_z) has been suppressed. Equation (1.31) implies that such a beamformer passes plane waves impinging from the look direction without attenuation or amplification. This is achieved by time-aligning the signals reaching each element in the array, and then summing them together *coherently*. As we will learn in the sequel, more advanced designs maintain this distortionless constraint, while simultaneously attempting to combine the signal components due to interference in a destructive manner, such that they are suppressed in the final output of the beamformer.

Although simple to analyze, it is well known that a linear array does not provide the optimal placement of sensors. The results of this section indicate that designing a microphone array for a broadband signal such as human speech involves a careful trade-off between achieving sufficient directivity at low frequencies, and avoiding spatial aliasing at high frequencies; see, for example, Van Trees (2002, §3.9.2) and Gazor and Grenier (1995). We will encounter these issues again in considering the design of spherical microphone arrays.

1.3 Conventional Adaptive Beamforming Algorithms

In this section, we discuss adaptive beamforming algorithms. In addition to passing a desired signal undistorted through the processing chain, such algorithms suppress unwanted noise, reverberation, or overlapping speech emanating from other directions. Hence, they are potentially far more effective at enhancing the desired signal than any fixed beamformer design.

1.3.1 Minimum Variance Distortionless Response Beamformer

The delay-and-sum beamformer can emphasize a wave emanating from a desired or look direction, and to some degree suppress waves impinging from other directions. As it is a *fixed* design, however, it does not provide optimal suppression for strong, coherent sources of interference. In contrast, the *adaptive beamformers* can effectively place a null on any interference by controlling the sidelobe structure of the beampattern, which is achieved by minimizing the variance of beamformer's outputs while maintaining a distortionless constraint in the look direction. This section describes one of the most basic adaptive beamforming methods, the *minimum variance distortionless response* (MVDR) beamformer. The MVDR beamformer is based on the use of *second-order statistics* (SOS); i.e., it requires only the knowledge of the covariance or *spatial spectral matrix* of the inputs to the microphone array.

For reasons of computational efficiency, modern adaptive filtering or beamforming algorithms are usually implemented in the frequency or—better yet—subband domain (Haykin 2002, §7). Section 1.3.3 briefly presents subband analysis and synthesis. Here, we consider beamforming in the subband domain. Let us define the *subband domain snapshot* to an array of S discrete sensors as

$$\mathbf{X}(\omega) \triangleq \begin{bmatrix} X_0(\omega) & X_1(\omega) & \cdots & X_{S-1}(\omega) \end{bmatrix}^T, \tag{1.32}$$

where $X_s(\omega)$ is the subband component for sensor s. For present purposes, let us assume that the complete snapshot consists of the sum

$$\mathbf{X}(\omega) = \mathbf{F}(\omega) + \mathbf{N}(\omega), \tag{1.33}$$

where $\mathbf{F}(\omega)$ is the component contributed by the desired signal and $\mathbf{N}(\omega)$ is due to the ambient noise or interference. Let us define the desired signal $F(\omega)$ which we assume to be transmitted on a plane wave with wavenumber $\mathbf{k}_\mathrm{s}$ impinging on the sensor array. Letting $\mathbf{m}_s$ denote the position of the sensor s leads to the representation

$$\mathbf{F}(\omega) \triangleq F(\omega)\mathbf{v}_\mathbf{k}(\mathbf{k}_\mathrm{s}), \tag{1.34}$$

where the array manifold vector in this case is defined as

$$\mathbf{v}_\mathbf{k}(\mathbf{k}_\mathrm{s}) \triangleq \begin{bmatrix} e^{-\mathbf{k}_\mathrm{s}^T\mathbf{m}_0} & e^{-\mathbf{k}_\mathrm{s}^T\mathbf{m}_1} & \cdots & e^{-\mathbf{k}_\mathrm{s}^T\mathbf{m}_{S-1}} \end{bmatrix}^T. \tag{1.35}$$

Alternatively, we can express the array manifold vector as

$$\mathbf{v}(\omega) \triangleq \begin{bmatrix} e^{-i\omega\tau_0} & e^{-i\omega\tau_1} & \cdots & e^{-i\omega\tau_{S-1}} \end{bmatrix}, \tag{1.36}$$

where

$$\omega\tau_s = \mathbf{k}_\mathrm{s}^T\mathbf{m}_s \;\; \forall\, s = 0, 1, \ldots, S-1.$$

Equation (1.36) is actually a more general definition of the array manifold vector than (1.35), inasmuch as the former encompasses spherical as well as plane waves; it is only necessary to modify the way in which τ_s is calculated. The output of the beamformer can then be expressed as

$$Y(\omega) = \mathbf{w}^H(\omega)\mathbf{X}(\omega), \tag{1.37}$$

where $\mathbf{w}^H(\omega)$ are the frequency-dependent sensor weights.

In order to calculate the optimal MVDR sensor weights, the covariance matrix of the outputs of the array sensors must be known or estimated. Here we assume they are known such that

$$\mathbf{\Sigma}_\mathbf{X}(\omega) \triangleq \mathcal{E}\left\{\mathbf{X}(\omega)\mathbf{X}^H(\omega)\right\}, \tag{1.38}$$

where $\mathcal{E}\{-\}$ is the probabilistic expectation operator (Papoulis and Pillai 2002, §5-3). We then determine the optimum weight vector that minimizes the variance of the beamformer's outputs

$$\Sigma_Y^2 \triangleq \mathcal{E}\left\{|Y(\omega)|^2\right\} = \mathbf{w}^H(\omega)\mathbf{\Sigma}_\mathbf{X}(\omega)\mathbf{w}(\omega), \tag{1.39}$$

subject to the distortionless constraint (1.31). The well-known solution is the MVDR beamformer (Wölfel and McDonough 2009, §13.3.1). The weight vector of the MVDR beamformer can be expressed as

$$\mathbf{w}_{\text{MVDR}}^{H}(\omega) = \frac{\mathbf{v}^{H}(\omega)\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}(\omega)}{\mathbf{v}^{H}(\omega)\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}(\omega)\mathbf{v}(\omega)}. \tag{1.40}$$

In practice, acoustic beamforming applications update the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}}$ only during periods of inactivity of the desired source in order to avoid cancellation of the desired signal, which is known as *signal cancellation* (Widrow et al. 1982). Van Trees (2002, §6.2.4) refers to the beamformer that uses the entire input for computation of the covariance matrix as the *minimum power distortionless response* (MPDR) beamformer, although both beamformers are commonly referred to as MVDR beamformers in the literature.

In order to avoid excessively large sidelobes in the beampattern, small weights are typically added to the main diagonal of $\boldsymbol{\Sigma}_{\mathbf{X}}$, which is known as *diagonal loading* (Wölfel and McDonough 2009, §13.3.7). Letting σ_{d}^2 denote the amount of diagonal loading, the weight vector of the MVDR beamformer can be written as

$$\mathbf{w}_{\text{MVDR}}^{H} = \frac{\mathbf{v}^{H}\left(\boldsymbol{\Sigma}_{\mathbf{X}} + \sigma_{\text{d}}^2\mathbf{I}\right)^{-1}}{\mathbf{v}^{H}\left(\boldsymbol{\Sigma}_{\mathbf{X}} + \sigma_{\text{d}}^2\mathbf{I}\right)^{-1}\mathbf{v}}, \tag{1.41}$$

where the frequency ω is omitted here for the sake of clarity.

Figure 1.9 shows the beampatterns of the MVDR beamformer constructed from the linear array with twenty equally-spaced sensors and an intersensor spacing of $d = \lambda/2$. In Figure 1.9, the diagonal loading is $\sigma_{\text{d}}^2 = 0.01$, and the look direction is set as $u = 0$; two interference signals are assumed to come from $u = -0.3$ and $u = 0.3$ as indicated with dotted lines. It is clear from the figure that the MVDR beamformer can maintain unity gain in the look direction at $u = 0$, while placing deep nulls on the directions of arrival of the interference at $u = \pm 0.3$.

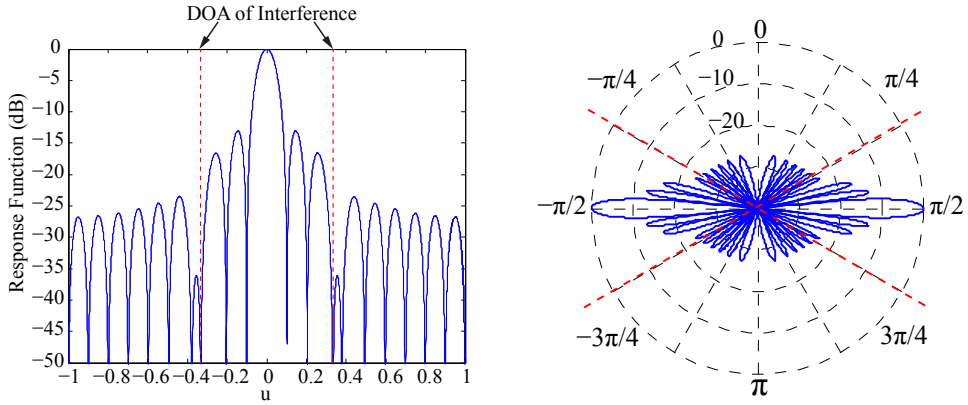
1.3.2 Noise Field Models

Although the MVDR beamformer can effectively suppress the interference signals by computing the noise covariance matrix from actual observations, it is often better to use a theoretical noise field model in practice. Two models that appear frequently in the literature are the incoherent and diffuse noise models.

In the case that a noise field is spatially uncorrelated (incoherent), the correlation of noise signals received at microphones at any given spatial location is zero. It was shown in Brandstein and Ward (2000, §4) that, under that condition, the noise covariance matrix becomes an identity matrix, that is, $\boldsymbol{\Sigma}_{\mathbf{X}}(\omega) = \mathbf{I}$. In that case, the MVDR solution for the sensor weights becomes equivalent to those of the delay-and-sum beamformer. The incoherent noise model is often appropriate when the distance between microphones is large and there are no coherent noise sources.

If the sensors of an array receive plane wave noise signals uniformly distributed on the surface of a sphere with random phase, a *spherically isotropic noise field* results. In this case, the components of the noise covariance matrix can be expressed as

$$\Sigma_{\mathbf{N}_{s,s'}}(\omega) = \text{sinc}\left(\frac{\omega d_{s,s'}}{c}\right), \tag{1.42}$$



a) Beampattern in Cartesian coordinates b) Polar plot of the beampattern

Figure 1.9 Beampatterns of the MVDR beamformer with $N = 20$ sensors and $d/\lambda = 1/2$ in the case of the look direction of $u = 0$ for two interference signals.

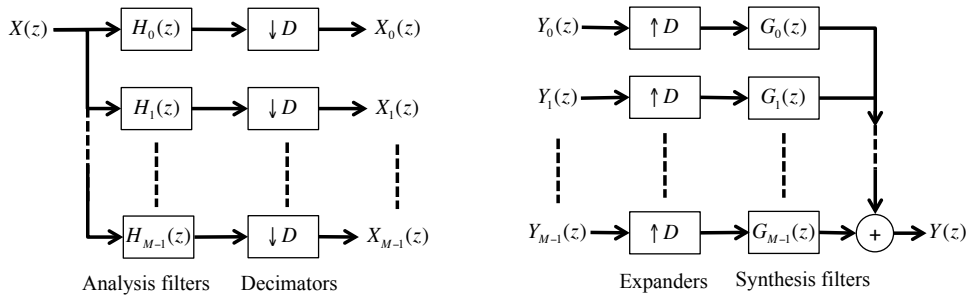
where $d_{s,s'}$ is the distance between microphones s and s'. When the weights (1.40) of the MVDR beamformer are estimated based on (1.42), the *superdirective beamformer* is obtained (Wölfel and McDonough 2009, §13.3.4). Another theoretical noise model frequently used in acoustic beamforming is the *cylindrically isotropic noise field*, which yields a sensor covariance matrix of

$$\Sigma_{\mathbf{N}_{s,s'}}(\omega) = J_0\left(\frac{\omega\, d_{s,s'}}{c}\right), \tag{1.43}$$

where J_0 is the *cylindrical Bessel function* of order zero (Olver and Maximon 2010, §10.2). The cylindrically isotropic noise field is said to be a good approximation for *babble noise* (Bitzer and Simmer 2001, §2.3.3).

1.3.3 Subband Analysis and Synthesis

Because of its computational efficiency, adaptive filtering and beamforming operations are often performed in the *frequency* or *subband* domain (Wölfel and McDonough 2009, §11), which provides additional advantages in terms of speed of convergence. Frequency domain analysis is typically performed by applying a windowing sequence $w[n]$, such as the *Hamming window*, to isolate a segment of the input, then performing a *discrete Fourier transform* (DFT) to this windowed sequence. This is equivalent to calculating the *short time Fourier transform* (STFT) of the segment (Oppenheim et al. 2009, §10.3). In principal, the same steps are also used for subband analysis. In the latter, however, if M subbands are to be used for analysis, the length of the window is typically mM for some integer $m > 1$, which implies that the windowed signal must be time aliased. The advantage afforded by the longer window is that the stop band suppression can be much greater than that achieved by frequency domain analysis (Wölfel and McDonough 2009, §11.8). This is a desirable characteristic for both adaptive filtering and beamforming as the outputs of all subbands can be treated as



a) Analysis processing for the input of the sth channel

b) Synthesis processing for the beamformer's output

Figure 1.10 Schematic of subband processing.

statistically independent; this independence is violated if there is significant spectral overlap between adjacent subbands. Moreover, the design of a subband analysis bank can be paired with that of a subband synthesis bank such that the combination is able to reconstruct the original input signal to arbitrary accuracy; i.e., it is able to achieve *perfect reconstruction*. Subband analysis and synthesis filter banks that are optimally-suited to adaptive filtering and beamforming applications achieve perfect reconstruction through *oversampling* rather than *aliasing cancellation* (Wölfel and McDonough 2009, §11).

A system for performing subband analysis and synthesis is shown in Figure 1.10. The set of transfer functions $\{H_m(z)\}$ comprises the *analysis filter bank*, which splits the input $x[n]$ into M subband signals $\{X_m[n]\}_{m=0}^{M-1}$. The set $\{G_m(z)\}$ of transfer functions comprises the *synthesis filter bank*, which recombines the M subband signals $\{Y_m[n]\}_{m=0}^{M-1}$ into a single output $\hat{x}[n]$. Each $Y_m[n]$ is obtained by multiplying $X_m[n]$ with a complex constant, which is determined with an adaptive filtering or beamforming operation. In the former case, there are a single analysis bank and a single synthesis bank; in the latter, there is one analysis bank for each element in a sensor array, but only a single synthesis bank as all the samples for a given subband are combined with a beamforming operation prior to synthesis.

A common class of filter banks is that wherein the impulse response of each filter is obtained by modulating a prototype impulse response $h_0[n]$ according to

$$h_m[n] = h_0[n]\, e^{j2\pi nm/M} \;\forall\; m = 0, \ldots, M-1, \tag{1.44}$$

which implies that the impulse responses for all the filters in the bank are obtained from a single prototype. The processes of windowing and filtering are then equivalent provided that $h_0[n] = w[-n]$. Applying the z-transform to both sides of (1.44), we obtain

$$H_m(z) \triangleq H_0(zW_M^m), \tag{1.45}$$

where $W_M = e^{-j2\pi/M}$ is the Mth root of unity. Equation (1.45) implies that $H_m(e^{j\omega})$ is a *shifted version* of the frequency response of $H_0(e^{j\omega})$ according to

$$H_m(e^{j\omega}) = H_0(e^{j(\omega - 2\pi m)/M}). \tag{1.46}$$

Similarly, for the synthesis bank, the impulse responses of the individual filters are related by

$$g_m[n] = g_0[n]\, e^{j2\pi nm/M} \;\forall\; m = 0, \ldots, M-1,$$

so that we can write

$$G_m(z) \triangleq G_0(zW_M^m). \tag{1.47}$$

We will now introduce two important operations in the filter bank system, *decimation* and *expansion*. Figure 1.10 also illustrates two corresponding blocks referred to as the D-fold *decimator* and the D-fold *expander*. The D-fold decimator with input $x[n]$ produces output

$$x_{\mathrm{D}}[n] = x[nD] \tag{1.48}$$

for integer D. In the frequency domain, the output of the decimator can be written as

$$X_{\mathrm{D}}(e^{j\omega}) = \frac{1}{D}\sum_{k=0}^{D-1} X(e^{j(\omega-2\pi k)/D}); \tag{1.49}$$

see Vaidyanathan (1993, §4.1). From (1.49), the operation of the decimator can be interpreted as follows: (1) stretch the input spectrum $X(e^{j\omega})$ by a factor of D in order to form $X(e^{j\omega/D})$, (2) create $D-1$ copies of the stretched spectrum by shifting it with an amount of 2π, (3) sum the original spectrum and all the stretched versions together and (4) divide it by D. Normally, each stretched version is overlapped with the other shifted copies. The effect of the overlap is known as *frequency aliasing*. In order to control such aliasing, the decimation factor D is set according to $D = M/2^r$ for some integer $r > 1$, which implies that the subbands are *oversampled*.

The D-fold expander takes input $y[n]$ and interpolates as

$$y_{\mathrm{E}}[n] = \begin{cases} y[n/D] & \text{if } n \text{ is an integer-multiple of } D, \\ 0 & \text{otherwise,} \end{cases} \tag{1.50}$$

Based on (1.50), we can write

$$Y_E(z) = \sum_{n=-\infty}^{\infty} y_E[n]\, z^{-n} = \sum_{k=-\infty}^{\infty} y_E[kD]\, z^{-kD} = \sum_{k=-\infty}^{\infty} y[k]\, z^{-kD}. \tag{1.51}$$

Upon setting $z = e^{j\omega}$ for the last equality, we have

$$Y_E(e^{j\omega}) = Y(e^{j\omega D}). \tag{1.52}$$

It is clear from (1.52) that the expander scales the frequency axis, which creates images of the compressed spectrum of $Y(e^{j\omega})$; this is known as *imaging*.

The filter bank obtained in the fashion described above is known as a *uniform DFT filter bank*. The uniform DFT analysis and synthesis filter banks are typically implemented in *polyphase* form in order to achieve maximal computational efficiency (Wölfel and McDonough 2009, §11). The task of designing the uniform DFT filter bank devolves to that of designing the analysis and synthesis prototypes $h_0[n]$ and $g_0[n]$, respectively.

In the class of cosine modulated filter banks, perfect reconstruction is achieved through aliasing cancellation (Vaidyanathan 1993, §5.6). During adaptive filtering or beamforming,

however, the perfect reconstruction property can be destroyed as arbitrary magnitude scalings and phase shifts are applied to the subband samples. De Haan et al. (2003) abandoned aliasing cancellation and designed analysis and synthesis prototypes based on minimization of the individual aliasing components for each subband. De Haan et al. also demonstrated that adaptive beamforming with their filter banks provides superior speech enhancement due to better suppression of aliasing effects; the latter can be further suppressed by imposing the *Nyquist(M)* constraint on the filter bank prototypes (Kumatani et al. 2008b).

Use of a digital filter bank requires that the array manifold vector (1.36) be redefined as

$$\mathbf{v}(\omega_m) \triangleq \left[e^{-i\omega_m \tau_0 f_s} \quad e^{-i\omega_m \tau_1 f_s} \quad \dots \quad e^{-i\omega_m \tau_{S-1} f_s}\right], \tag{1.53}$$

where f_s is the digital sampling frequency.

1.3.4 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 1.5.

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 (1.30) and (1.40), respectively.

As in Section 1.3.1, 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

$$\text{SNR}_{\text{in}}(\omega) \triangleq \frac{\Sigma_F(\omega)}{\Sigma_N(\omega)}, \tag{1.54}$$

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

$$Y(\omega) = \mathbf{w}^H(\omega)\,\mathbf{X}(\omega) = Y_F(\omega) + Y_N(\omega), \tag{1.55}$$

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

$$\Sigma_Y(\omega) = \mathcal{E}\{|Y(\omega)|^2\} = \Sigma_{Y_F}(\omega) + \Sigma_{Y_N}(\omega), \tag{1.56}$$

where

$$\Sigma_{Y_F}(\omega) \triangleq \mathbf{w}^H(\omega)\,\mathbf{\Sigma_F}(\omega)\,\mathbf{w}(\omega) \tag{1.57}$$

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

$$\Sigma_{Y_N}(\omega) \triangleq \mathbf{w}^H(\omega)\,\mathbf{\Sigma_N}(\omega)\,\mathbf{w}(\omega) \tag{1.58}$$

is the variance of the noise component. Expressing the snapshot of the desired signal once more as in (1.32), we find that the spatial spectral matrix $\mathbf{F}(\omega)$ of the desired signal can be written as

$$\mathbf{\Sigma_F}(\omega) = \Sigma_F(\omega)\,\mathbf{v_k}(\mathbf{k}_\mathrm{s})\,\mathbf{v_k^H}(\mathbf{k}_\mathrm{s}). \tag{1.59}$$

Substituting (1.59) into (1.57), we can calculate the variance of the output signal spectrum as

$$\Sigma_{Y_F}(\omega) = \mathbf{w}^H(\omega)\,\mathbf{v_k}(\mathbf{k}_\mathrm{s})\,\Sigma_F(\omega)\,\mathbf{v_k^H}(\mathbf{k}_\mathrm{s})\,\mathbf{w}(\omega). \tag{1.60}$$

If we now assume that $\mathbf{w}(\omega)$ satisfies the distortionless constraint (1.31), then (1.60) reduces to

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

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

Substituting (1.30) into (1.58) it follows that the noise component present at the output of the DSB is given by

$$\Sigma_{Y_N}(\omega) = \frac{1}{N^2}\mathbf{v_k^H}(\mathbf{k}_\mathrm{s})\,\mathbf{\Sigma_N}(\omega)\,\mathbf{v_k}(\mathbf{k}_\mathrm{s}) \tag{1.61}$$

$$= \frac{1}{N^2}\mathbf{v_k^H}(\mathbf{k}_\mathrm{s})\boldsymbol{\rho}_\mathbf{N}(\omega)\mathbf{v_k}(\mathbf{k}_\mathrm{s})\Sigma_N(\omega), \tag{1.62}$$

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

$$\mathbf{\Sigma_N}(\omega) \triangleq \Sigma_N(\omega)\,\boldsymbol{\rho}_\mathbf{N}(\omega). \tag{1.63}$$

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

$$\mathrm{SNR}_{\mathrm{out}}(\omega) \triangleq \frac{\Sigma_{Y_F}(\omega)}{\Sigma_{Y_N}(\omega)} = \frac{\Sigma_F(\omega)}{\mathbf{w}^H(\omega)\,\mathbf{\Sigma_N}(\omega)\mathbf{w}(\omega)}. \tag{1.64}$$

Then based on (1.54) and (1.64), we can calculate the array gain of the DSB as

$$A_{\mathrm{dsb}}(\omega,\mathbf{k}_\mathrm{s}) \triangleq \frac{\Sigma_{Y_F}(\omega)}{\Sigma_{Y_N}(\omega)}\Big/\frac{\Sigma_F(\omega)}{\Sigma_N(\omega)} = \frac{N^2}{\mathbf{v_k^H}(\mathbf{k}_\mathrm{s})\,\boldsymbol{\rho}_\mathbf{N}(\omega)\,\mathbf{v_k}(\mathbf{k}_\mathrm{s})}. \tag{1.65}$$

Repeating the foregoing analysis for the MVDR beamformer (1.40), we arrive at

$$A_{\mathrm{mvdr}}(\omega,\mathbf{k}_\mathrm{s}) = \mathbf{v_k^H}(\mathbf{k}_\mathrm{s})\,\boldsymbol{\rho}_\mathbf{N}^{-1}(\omega)\,\mathbf{v_k}(\mathbf{k}_\mathrm{s}). \tag{1.66}$$

If noise at all sensors are spatially uncorrelated, then $\boldsymbol{\rho}_\mathbf{N}(\omega)$ is the identity matrix and the MVDR beamformer reduces to the DSB. From (1.65) and (1.66), it can be seen that in this case, the array again is

$$A_{\mathrm{mvdr}}(\omega,\mathbf{k}_\mathrm{s}) = A_{\mathrm{dsb}}(\omega,\mathbf{k}_\mathrm{s}) = N. \tag{1.67}$$

In all other cases,

$$A_{\text{mvdr}}(\omega, \mathbf{k}_{\text{s}}) > A_{\text{dsb}}(\omega, \mathbf{k}_{\text{s}}). \tag{1.68}$$

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 (Van Trees 2002, §6.2.3). Secondly, and more importantly, by placing a Wiener filter (Haykin 2002, §2.2) on the output of the MVDR beamformer, the *minimum mean-square error* (MMSE) beamformer is obtained (Van Trees 2002, §6.2.2). Such *post-filters* are important because it has been shown that they can yield significant reductions in error rate (McCowan and Bourlard 2003; McDonough et al. 2007). If only a single subband is considered, the MVDR beamformer without modification will uniformly provide the highest SNR, as indicated by (1.68), and hence the highest array gain; we will return to this point in Section 1.5.

White Noise Gain

The *white noise gain* (WNG) is by definition (Cox et al. 1987)

$$G_{\text{w}}(\omega) \triangleq \frac{\left|\mathbf{w}^H(\omega)\,\mathbf{v}(\mathbf{k}_{\text{s}})\right|^2}{\mathbf{w}^H(\omega)\,\mathbf{w}(\omega)}. \tag{1.69}$$

The numerator of (1.69), which will be unity for any beamformer satisfying the distortionless constraint (1.31), 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 (1955) 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 (Van Trees 2002, §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 1.5.

Directivity Index

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

$$P(\theta, \phi) \triangleq \left|B(\theta, \phi)\right|^2, \tag{1.70}$$

where $B(\theta, \phi)$ is the beampattern described in Section 1.2 as a function of the spherical coordinates $\Omega \triangleq (\theta, \phi)$; see Figure 1.3. 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 (Van Trees 2002, §2.6.1)

$$D(\omega) \triangleq \frac{4\pi P(\theta_0, \phi_0)}{\displaystyle\int_{\Omega_{\text{sph}}} P(\theta, \phi)\, d\Omega}, \tag{1.71}$$

where Ω_{sph} represents the surface of a sphere with differential area $d\Omega$; we will consider such spherical integrals in detail in Sections 1.4 and 1.5.

Assuming that the beamforming coefficients satisfy the distortionless constraint (1.31) implies $P(\Omega_0) = 1$ such that (1.71) can be simplified and expressed in decibels as the *directivity index*

$$\begin{aligned} \text{DI} &\triangleq -10\log_{10}\left[\frac{1}{4\pi}\int_{\Omega} P(\theta,\phi)\,d\Omega\right] \\ &= -10\log_{10}\left[\frac{1}{4\pi}\int_{0}^{2\pi}\int_{0}^{\pi} P(\theta,\omega)\sin\theta d\theta d\phi\right]. \end{aligned} \tag{1.72}$$

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 (1.42); see Bitzer and Simmer (2001). Under this definition, the directivity index can be expressed as

$$\text{DI} \triangleq -10\log_{10}\frac{\left|\mathbf{w}^H\mathbf{v}(\mathbf{k}_S)\right|^2}{\mathbf{w}^H\boldsymbol{\Gamma}_{SI}\mathbf{w}}. \tag{1.73}$$

The superdirective beamformer mentioned in Section 1.3.2 will uniformly provide the highest directivity index, although this may not be the case when the covariance matrix (1.42) is diagonally loaded to achieve greater robustness. We will return to this point in Section 1.5.

1.3.5 Generalized Sidelobe Canceller Implementation

The MVDR beamformer can be also realized with a *generalized sidelobe canceller* (GSC). Figure 1.11 illustrates the beamformer in GSC configuration.

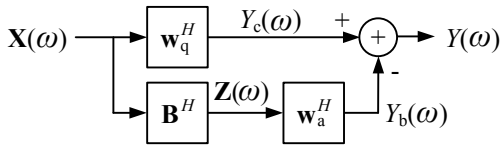


Figure 1.11 Generalized sidelobe canceller.

From here, we suppress the frequency index ω for the sake of convenience. The weights of the GSC beamformer consists of three components, the quiescent weight vector $\mathbf{w}_q$, the blocking matrix $\mathbf{B}$ and the active weight vector $\mathbf{w}_a$. The output of the beamformer at frame k for a given subband can be expressed as

$$Y(k) = \left(\mathbf{w}_q - \mathbf{B}\mathbf{w}_a\right)^H \mathbf{X}(k). \tag{1.74}$$

In keeping with the GSC formalism , $\mathbf{w}_{\mathrm{q}}$ is chosen to give unity gain in the desired look direction (Wölfel and McDonough 2009, §13.3.7). The blocking matrix is chosen to be orthogonal to the quiescent vector such that $\mathbf{B}^H\mathbf{w}_{\mathrm{q}} = \mathbf{0}$. The blocking matrix can be, for example, calculated with an orthogonalization technique such as the modified Gram-Schmidt method, QR decomposition or singular value decomposition (SVD) Golub and Van Loan (1996). The orthogonality implies that the distortionless constraint will be satisfied for any choice of $\mathbf{w}_{\mathrm{a}}$. Note that the blocking matrix is not unique.

In the case that the position of a sound source is static, the active weight vector is typically adjusted so that the variance of the GSC beamformer's outputs is minimized. Without diagonal loading, the solution of the active weight vector can be expressed as

$$\mathbf{w}_{\mathrm{a}}{}^H = \mathbf{w}_{\mathrm{q}}{}^H\mathbf{\Sigma_X}\mathbf{B}\left(\mathbf{B}^H\mathbf{\Sigma_X}\mathbf{B}\right)^{-1}, \tag{1.75}$$

where $\mathbf{\Sigma_X}$ is the covariance matrix of the input vectors.

The interference signal can be effectively suppressed based on equations (1.40), (1.41) or (1.75). However, they are not suitable for the online implementation because it assumes that the second order statistics, $\mathbf{\Sigma_X}$, are known from a sufficient amount of data. It is preferably updated on a sample-by-sample basis. In the next section, the online algorithm of the conventional GSC beamformer will be described.

1.3.6 Recursive Implementation of the GSC

In many applications, the active weight vector of the GSC beamformer is updated at each frame by using a *recursive least squares* (RLS) method or a *least mean square* (LMS) algorithm. Here, we review, without formal proof, the RLS algorithm and briefly comment on the differences between the RLS and LMS algorithms. Further details can be found in (Van Trees 2002, §7.4) and (Van Trees 2002, §7.6), respectively.

In order to recursively update the active weight vector at frame T while retaining the information from frames $t = 0, 1, \ldots, T-1$, we first introduce *forgetting factor* $0 < \mu < 1$ and define the exponentially-weighted spatial spectral matrix,

$$\mathbf{\Phi}_X(T) \triangleq \sum_{t=1}^{T} \mu^{T-t}\mathbf{X}(t)\mathbf{X}^H(t). \tag{1.76}$$

Similarly, the exponentially-weighted spatial spectral matrix of the output $\mathbf{Z}(t)$ of the blocking matrix is

$$\mathbf{\Phi}_Z(T) \triangleq \sum_{t=1}^{T} \mu^{T-t}\mathbf{Z}(t)\mathbf{Z}^H(t) = \mathbf{B}^H\mathbf{\Phi}_X(T)\mathbf{B}. \tag{1.77}$$

Finally, the cross-correlation between the blocking matrix's output $\mathbf{Z}(t)$ and the quiescent vector's output $Y_c(t)$ is given by

$$\mathbf{\Phi}_{ZY_c^*}(T) \triangleq \sum_{t=1}^{T} \mu^{T-t}\mathbf{Z}(t)Y_c^*(t) = \mathbf{B}^H\mathbf{\Phi}_X(T)\mathbf{w}_{\mathrm{q}}. \tag{1.78}$$

Now let us define

$$\mathbf{P}_Z(T) = \mathbf{\Phi}_Z^{-1}(T) \text{ and} \tag{1.79}$$

$$\mathbf{g}_Z(T) = \frac{\mu^{-1}\mathbf{P}_Z(T-1)\mathbf{Z}(T)}{1+\mu^{-1}\mathbf{Z}^H(T)\mathbf{P}_Z(T-1)\mathbf{Z}(T)}. \tag{1.80}$$

The notations are deliberately chosen by considering the relationship between the RLS algorithm and the Kalman filter (Haykin 2002, §10.8). In this case, the well-known Riccati equation can be expressed as

$$\mathbf{P}_Z(T) = \mu^{-1}\mathbf{P}_Z(T-1) - \mu^{-1}\mathbf{g}_Z(T)\mathbf{Z}^H(T)\mathbf{P}_Z(T-1). \tag{1.81}$$

Post-multiplying both sides of (1.81) by $\mathbf{Z}(T)$ and substituting the resulting equality into (1.80), we find the gain vector

$$\mathbf{g}_Z(T) = \mathbf{P}_Z(T)\mathbf{Z}(T). \tag{1.82}$$

The goal of the RLS method is to minimize a weighted sum of array outputs $Y(t)$ defined as

$$\Phi_Y(T) \triangleq \sum_{t=1}^{T} \mu^{T-t}|Y(t)|^2. \tag{1.83}$$

Note that the importance of the past outputs decreases exponentially with time T. Upon taking the derivative of (1.83) with respect to $\mathbf{w}_{\mathrm{a}}(T)$ and setting the result to zero, we find

$$\mathbf{w}_{\mathrm{a}}(T) = \mathbf{\Phi}_Z^{-1}(T)\mathbf{\Phi}_{ZY_c^*}(T) = \mathbf{P}_Z(T)\mathbf{\Phi}_{ZY_c^*}(T). \tag{1.84}$$

Substituting (1.78) and (1.81) into (1.84), we obtain

$$\mathbf{w}_{\mathrm{a}}(T) = \mathbf{w}_{\mathrm{a}}(T-1) + \mathbf{g}_Z(T)e_p^*(T). \tag{1.85}$$

where

$$e_p(T) = Y_c(T) - {\mathbf{w}_{\mathrm{a}}}^H(T-1)\mathbf{Z}(T). \tag{1.86}$$

At each frame, we can add the diagonal component σ_z^2 to $\mathbf{Z}(t)\mathbf{Z}^H(t)$. The RLS update formula with diagonal loading can be then written as

$$\mathbf{w}_{\mathrm{a}}(T) = \left[\mathbf{I} - \sigma_z^2\mathbf{P}_Z(T)\right]\mathbf{w}_{\mathrm{a}}(T-1) + \mathbf{g}_Z(T)e_p^*(T), \tag{1.87}$$

Notice that (1.86) does not directly load the diagonal component of the sample spectral matrix in contrast to (1.41).

The update algorithms of (1.86) and (1.87) are suitable for on-line operation because the active weight vector can be adapted with the instantaneous input vector.

The difference between the RLS and LMS implementations is the step size parameter for the update formula. The LMS algorithm has a choice to adjust the step size parameter whereas the GSC-RLS algorithm uses $\mathbf{\Phi}_Z^{-1}$ as the step size. Multiplying the gradient with $\mathbf{\Phi}_Z^{-1}$ enables each component of the active weight vector to converge at the same rate. The disadvantage would be additional computation although it is normally trivial.

1.3.7 Other Conventional GSC Beamformers

In theory, the conventional MVDR beamfomers described above can eliminate interfering signals. In practice, however, they are prone to the signal cancellation problem whenever there is an interfering signal that is correlated with the desired signal. In realistic acoustic environments, interfering signals are highly correlated with the desired signal, as the latter is reflected from hard surfaces such as walls and tables and thereafter impinges on the sensor array from directions that are distinct from the look direction. Beam steering errors as well as magnitude and phase errors in the frequency responses of the individual sensors in an array can also cause signal cancellation to occur.

To avoid the signal cancellation, many algorithms have been developed. Those approaches can be classified into the following categories:

1. Updating the active weight vector only when noise signals are dominant (Cohen et al. 2003; Herbordt and Kellermann 2003; Nordholm et al. 1993);
2. Constraining the update formula for the active weight vector (Claesson and Nordholm 1992; Hoshuyama et al. 1999; Nordebo et al. 1994);
3. Blocking the leakage of desired signal components into the sidelobe canceller by designing the blocking matrix (Herbordt and Kellermann 2002; Herbordt et al. 2007; Hoshuyama et al. 1999; Warsitz et al. 2008);
4. Taking speech distortion due to the the leakage of a desired signal into account using a multi-channel Wiener filter which aims at minimizing a weighted sum of residual noise and speech distortion terms (Doclo et al. 2007);
5. Using acoustic transfer functions from the desired source to each sensor in an array instead of merely compensating for time delays (Cohen et al. 2003; Gannot and Cohen 2004; Sharon Gannot et al. 2001; Warsitz et al. 2008).

As mentioned before, the algorithms discussed in this section minimize nearly the same criterion based on the second-order statistics (SOS) such as the variance of the beamformer's outputs. In the following section, we describe beamforming algorithms which adjust the active weight vector based on higher-order statistics (HOS); these algorithms have appeared more recently in the literature.

1.3.8 Beamforming based on Higher-Order Statistics

A multidimensional Gaussian pdf is completely characterized once its mean vector and covariance matrix are known. Hence, speech enhancement techniques that assume—either implicitly or explicitly—that speech is a Gaussian random process are said to be second-order methods. For any non-Gaussian pdf on the other hand, the higher-order statistical moments have a great deal of influence on the fine structure of the pdf. Thus, enhancement techniques that take into account the deviation from Gaussianity inherent in human speech are said to based on *higher-order statistics* (HOS). Such techniques are the subject of this section.

Statistical Characteristics of Human Speech

In order to avoid the signal cancellation problem, HOS recently have been introduced to the field of acoustic beamforming. HOS have long been used in the field of *independent component analysis* (ICA) (Hyvärinen 1999).

The entire field of ICA is founded on the assumption that all signals of interest are not Gaussian-distributed (Hyvärinen and Oja 2000). Briefly, the reasoning is grounded on two points:

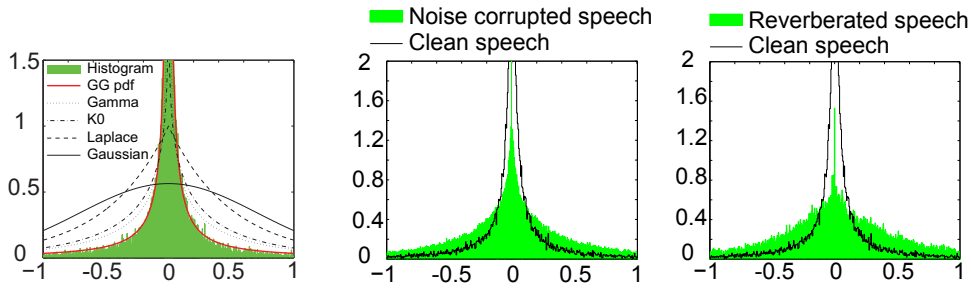
1. The *central limit theorem* states that the pdf of the sum of independent random variables (RVs) will approach Gaussian in the limit as more and more components are added, *regardless* of the pdfs of the individual components. This implies that the sum of several RVs will be closer to Gaussian than any of the individual components.
2. The *entropy* for a complex-valued RV Y is defined as

$$H(Y) \triangleq -\mathcal{E}\left\{\log p_Y(v)\right\} = -\int p_Y(v) \log p_Y(v) dv, \tag{1.88}$$

where $p_Y(.)$ is the pdf of Y. The integral form of the entropy for the continuous RVs with the pdfs is referred to the *differential entropy*, or simply *entropy*, and distinguished from an ensemble average for samples. The entropy is the basic measure of information in *information theory* (Gallager 1968). It is well known that a Gaussian RV has the highest entropy of all RVs with a given variance (Gallager 1968, Thm. 7.4.1), which also holds for complex Gaussian RVs (Neeser and Massey 1993, Thm. 2). Hence, a Gaussian RV is, in some sense, the least *predictable* of all RVs. Information-bearing signals, on the other hand, are redundant and thus contain structure that makes them more predictable than Gaussian RVs.

These points suggest that one must look for the *least* Gaussian RV in order to obtain the information-bearing signals. The fact that the pdf of speech is super-Gaussian has often been reported in the literature (Erkelens et al. 2007; Kumatani et al. 2007; Martin 2005). Noise, on the other hand, is more nearly Gaussian-distributed. In fact, the pdf of the sum of several super-Gaussian RVs becomes closer to Gaussian. Thus, a mixture consisting of a desired signal and several interfering signals can be expected to be nearly Gaussian-distributed.

Figure 1.12 a) shows a histogram of the real parts of subband samples of speech at frequency 800 Hz. To generate these histograms, the authors used 43.9 minutes of clean speech recorded with a close-talking microphone (CTM) from the development set of the Speech Separation Challenge, Part 2 (SSC2) (Lincoln et al. 2005). The Gaussian, Laplace, K_0 , Γ (Kumatani et al. 2007), and GG pdfs (Kumatani et al. 2008a) are also shown in Figure 1.12 a). In Figure 1.12 a), the parameters of each pdf were estimated from training data based on the maximum likelihood criterion. It is clear from Figure 1.12 that the distribution of clean speech is not Gaussian but super-Gaussian. Figure 1.12 a) suggests that the GG pdf is well suited for modeling subband samples of speech. From Figure 1.12 a), it is also clear that the Laplace, K_0, Γ, and GG pdfs exhibit the spikey and heavy-tailed characteristics. Super-Gaussian pdfs have a sharp concentration of probability mass at the mean, relatively little probability mass as compared with the Gaussian at intermediate values of the argument, and a relatively large amount of probability mass in the tail; i.e., far from the mean.



a) Super-Gaussian pdf b) Noise-corrupted speech c) Reverberated speech

Figure 1.12 Histograms of real parts of subband frequency components of clean speech and a) pdfs, b) noise-corrupted speech and c) reverberated speech.

Figure 1.12 b) shows histograms of real parts of subband components calculated from clean speech and noise-corrupted speech. The figure indicates that the pdf of the noise-corrupted signal, which is in fact the sum of the speech and noise signals, is closer to Gaussian than that of clean speech.

Figure 1.12 c) shows histograms of clean speech and reverberant speech in the subband domain. In order to produce the reverberant speech, a clean speech signal was convolved with an impulse response measured in a room; see Lincoln et al. (2005) for the configuration of the room. We can observe from Figure 1.12 c) that the pdf of reverberated speech is also closer to Gaussian than the original clean speech.

Maximum Kurtosis Beamforming

The *excess kurtosis* or simply *kurtosis* of a RV Y with zero mean is defined as

$$\text{kurt}(Y) \triangleq \mathcal{E}\{|Y|^4\} - \beta(\mathcal{E}\{|Y|^2\})^2. \tag{1.89}$$

where β is a positive constant, which is typically set to three for kurtosis of real-valued RVs in order to ensure that the Gaussian pdf has zero kurtosis; pdfs with positive kurtosis are super-Gaussian, and those with negative kurtosis are sub-Gaussian.

As indicated in (1.89), the kurtosis measure considers not only the variance but also the fourth moment, a higher-order statistic. As mentioned previously, any Gaussian pdf is completely specified its mean and variance; the higher-order statistics are not required.

In practice, the kurtosis of T outputs $Y(t)$ from a beamformer can be measured by simply averaging samples according to

$$J_{\text{kurt}}(Y) \triangleq \frac{1}{T}\sum_{t=0}^{T-1}|Y(t)|^4 - \beta\left(\frac{1}{T}\sum_{t=0}^{T-1}|Y(t)|^2\right)^2, \tag{1.90}$$

where the frequency index ω is omitted for clarity. The kurtosis criterion does not require any explicit assumption as to the exact form of the pdf. Due to its simplicity, it is widely used as a measure of non-Gaussianity. As demonstrated in Kumatani et al. (2007, 2008a), maximizing the degree of super-Gaussianity yields an active weight vector $\mathbf{w}_{\text{a}}$ capable of canceling

interference—including incoherent noise that leaks through the sidelobes—without the signal cancellation problem encountered in conventional beamforming.

As discussed in Section 1.3.1, diagonal loading is typically used in beamforming with SOS in order to reduce the norm of the active weight vector and thereby improve robustness by inhibiting the formation of excessively large sidelobes (Wölfel and McDonough 2009, §13.3.8). Such a regularization term can be also applied to maximum kurtosis beamforming by defining the modified optimization criterion of (1.90) with a weight parameter α as

$$\mathcal{J}_{\text{kurt}}(Y;\alpha) \triangleq J_{\text{kurt}}(Y) - \alpha\|\mathbf{w}_{\text{a}}\|^2 \quad \alpha > 0. \tag{1.91}$$

In Kumatani et al. (2008a), the sensitivity of the weight parameter α was investigated in terms of speech recognition. The best recognition performance was obtained with $\alpha = 0.01$ although the effect was not significant.

Unfortunately, there is no closed-form solution for the active weight vector which provides the maximum kurtosis. Thus, we have to resort to numerical optimization algorithms such as gradient descent. Upon substituting (1.74) into (1.91) and taking the partial derivative with respect to $\mathbf{w}_{\text{a}}$, we obtain

$$\frac{\partial \mathcal{J}_{\text{kurt}}(Y;\alpha)}{\partial \mathbf{w}_{\text{a}}{}^*} = \frac{2}{T}\sum_{t=0}^{T-1}\left\{-|Y(t)|^2 + \beta\sigma_Y^2\right\}\mathbf{B}^H(t)\mathbf{X}(t)Y^*(t) - \alpha\mathbf{w}_{\text{a}}, \tag{1.92}$$

where σ_Y^2 is the variance of beamformer's outputs.

Equation (1.92) is sufficient to implement a numerical optimization algorithm based, for example, on the method of steepest descent (Bertsekas 1995, §1.6), whereby kurtosis of beamformer's outputs can be maximized. The norm of the active weight vector is usually normalized in addition to the regularization term because it tends to become large.

Maximum Negentropy Beamforming

Another criterion for measuring the degree of super-Gaussianity is negentropy. The negentropy of a complex-valued RV Y is defined as

$$\text{neg}(Y) \triangleq H(Y_{\text{gauss}}) - \beta H(Y) \tag{1.93}$$

where Y_{gauss} is a Gaussian variable with the same variance σ_Y^2 as Y, β is a positive constant for adjusting the equilibrium condition and normally set to unity for negentropy. The entropy of Y_{gauss} can be expressed as

$$H(Y_{\text{gauss}}) = \log\left|\sigma_Y^2\right| + (1 + \log\pi)\,. \tag{1.94}$$

Note that negentropy is non-negative, and zero if and only if Y has a Gaussian distribution. Clearly, it can measure how far the desired distribution is far from the Gaussian pdf. Computing the entropy of the super-Gaussian variables $H(Y)$ requires knowledge of their specific pdf. Thus it is important to find a family of pdfs capable of closely modeling the distributions of actual speech signals.

The value calculated for kurtosis, however, can be strongly influenced by a few samples with a low observation probability. Hyvärinen and Oja (2000) demonstrates that negentropy is generally more robust in the presence of outliers than kurtosis.

As shown in Figure 1.12 a), the distribution of the subbands of clean speech can be represented with the generalized Gaussian pdf . In the case that the complex-valued RV Y possesses circular symmetry, a complex-valued GG pdf can be expressed as

$$p_{\mathrm{GG}}(Y) = \frac{f}{2\pi B^2(f)\Gamma(2/f)\hat{\sigma}^2} \exp\left[- \left| \frac{Y}{\hat{\sigma} B(f)} \right|^f \right], \tag{1.95}$$

where

$$B(f) = \left[\frac{\Gamma(2/f)}{\Gamma(4/f)} \right]^{1/2} \text{ and} \tag{1.96}$$

$\Gamma(.)$ is the gamma function (Askey and Roy 2010, §5.2). Note that the GG with $f = 2$ corresponds to the Gaussian pdf, whereas the GG pdf converges to a uniform distribution in the case of $f \to +\infty$.

The parameters of the GG pdf can be, for example, estimated using the maximum likelihood (ML) criterion, as in Wölfel and McDonough (2009, §13.5.2) and Kumatani et al. (2010b). The shape parameters are estimated independently for each subband, as the optimal pdf is frequency-dependent.

In order to develop maximum negentropy beamforming, we compute an ensemble average of negative log-likelihoods instead of differential entropy. In this case, negentropy of T frames of output from the array can be calculated according to

$$J_{\mathrm{neg}}(Y) = -\frac{1}{T} \sum_{t=0}^{T-1} \log p_{\mathrm{gauss}}\left(Y(t)\right) + \beta \frac{1}{T} \sum_{t=0}^{T-1} \log p_{\mathrm{GG}}\left(Y(t)\right), \tag{1.97}$$

where $p_{\mathrm{gauss}}(.)$ is the complex Gaussian pdf. Similar to (1.91), a regularization term can be added to the empirical negentropy to provide the modified optimization criterion,

$$\mathcal{J}_{\mathrm{neg}}(Y;\alpha) = J_{\mathrm{neg}}(Y) - \alpha \|\mathbf{w}_{\mathrm{a}}\|^2 \quad \alpha > 0. \tag{1.98}$$

Upon substituting (1.74) into (1.98) and taking the partial derivative, we obtain

$$\frac{\partial \mathcal{J}_{\mathrm{neg}}(Y;\alpha)}{\partial \mathbf{w}_{\mathrm{a}}{}^*} = \frac{1}{T} \sum_{t=0}^{T-1} \left\{ -\frac{1}{\sigma_Y^2} + \beta \frac{f|Y(t)|^{f-2}}{2(B(f)\hat{\sigma})^f} \right\} \mathbf{B}^H(t)\mathbf{X}(t)Y^*(t) - \alpha \mathbf{w}_{\mathrm{a}}. \tag{1.99}$$

The HOS-based beamformers, maximum kurtosis and maximum negentropy beamformers, do not suffer from signal cancellation encountered in the SOS-based adaptive beamformers. Therefore, the active weight vector can be adapted even when the desired source is active. Indeed, Kumatani et al. (2007, 2008a) demonstrated through acoustic simulations that beamformers based on HOS can emphasize the desired signal by coherently adding its reflections after a suitable phase shift in the subband domain. Hence, adaptation of the active weight vector is best performed *while* the desired speaker is active.

1.3.9 Online Implementation

Adaptive beamforming algorithms require a certain amount of data for stable estimation of the active weight vector. In the case of HOS-based beamforming, this problem becomes significant because it entirely relies on numerical optimization algorithms.

In order to achieve efficient estimation, a subspace (eigenspace) filter (Van Trees 2002, §6.8) can be used as a pre-processing step for estimation of the active weight vector . Motivations behind this idea are to 1) reduce the dimensionality of the active weight vector , and 2) improve speech enhancement performance based on decomposition of the outputs of the blocking matrix into spatially correlated and ambient signal components. Such decomposition can be achieved by performing an eigendecomposition on the covariance matrix of blocking matrix's outputs. Then, we select the eigenvectors corresponding to the largest eigenvalues as the dominant modes (Van Trees 2002, §6.8.3). The dominant modes are associated with the spatially correlated signals and the other modes are averaged as a signal model of ambient noise. By doing so, we can readily subtract the averaged ambient noise component from the beamformer's output. Moreover, the reduction of the dimension of the active weight leads to computationally efficient and reliable estimation. Notice that we adjust the active weight vector based on the maximum kurtosis criterion in contrast to the normal dominant-mode rejection (DMR) beamformers (Van Trees 2002, §6.8.3). It is also worth noting that subspace filtering here is analogous to whitening used as a measure of pre-processing in the field of ICA (Hyvärinen and Oja 2000).

In the following sections, we first discuss the subspace method for maximum kurtosis beamforming and then describe its online implementation. In those sections, we omit the frequency index ω for the sake of convenience.

Subspace Method

In the case that there are neither steering errors nor mismatches between microphones, the blocking matrix's output $\mathbf{Z}(t)$ only contains the spatially correlated (coherent) interference and ambient (incoherent) noise signals. However, in the real environments, it also includes the desired signal components due to those errors as well as the reverberation effects.

Let us first denote the D spatially correlated signal components contained in the output of the $N \times (N-1)$ blocking matrix as

$$\mathbf{V}(t) = [V_0(t), \cdots, V_d(t), \cdots, V_{D-1}(t)]^T . \tag{1.100}$$

Then, the output of the blocking matrix can be expressed as

$$\mathbf{Z}(t) = \mathbf{A}\mathbf{V}(t) + \mathbf{N}(t) \tag{1.101}$$

where $\mathbf{A}$ and $\mathbf{N}(t)$ represent a mixing matrix and the ambient noise signals, respectively. The direct path from the desired source signal to each microphone is assumed to be excluded from $\mathbf{A}$ because of the distortionless constraint imposed with the blocking matrix. Thus, in the case that there is neither reverberation nor error such as a microphone array mismatch and steering error, $\mathbf{Z}$ consists of the interference signals only.

Assuming that $\mathbf{V}(t)$ and $\mathbf{N}(t)$ are uncorrelated, we can write the covariance matrix of $\mathbf{Z}$ as

$$\Sigma_{\mathbf{Z}} = \mathcal{E}\left[\mathbf{Z}(t)\mathbf{Z}^H(t)\right] = \mathbf{A}\Sigma_{\mathbf{V}}\mathbf{A}^H + \Sigma_{\mathbf{N}}, \tag{1.102}$$

where

$$\Sigma_{\mathbf{V}} = \mathcal{E}\left[\mathbf{V}(t)\mathbf{V}^H(t)\right] \text{ and } \Sigma_{\mathbf{N}} = \mathcal{E}\left[\mathbf{N}(t)\mathbf{N}^H(t)\right].$$

The subspace method seeks a set of D linearly independent vectors contained in the subspace, $\Re\{\mathbf{A}\}$, spanned by the column vectors of $\mathbf{A}$. The first step for obtaining such set

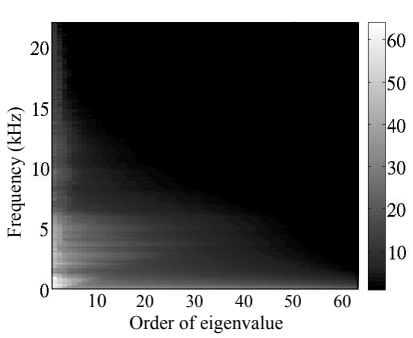


Figure 1.13 Three-dimensional representation of the eigenvalue distribution as a function of the order and frequencies.

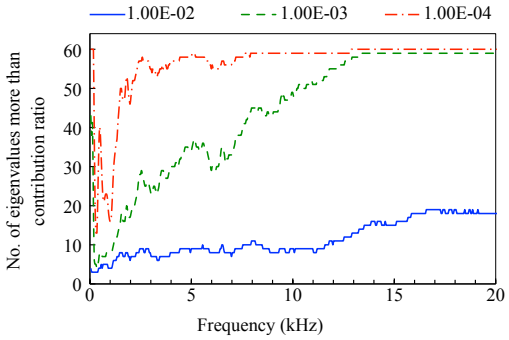


Figure 1.14 The average of the numbers of the contribution ratios that exceed thresholds 10^{-2} as a function of frequencies.

of the vectors is to solve the generalized eigenvalue (GE) decomposition problem as in Roy and Kailath (1989); Van Trees (2002, §6.8),

$$\Sigma_{\mathbf{Z}}\mathbf{E} = \Sigma_{\mathbf{N}}\mathbf{E}\mathbf{\Lambda},$$

where $\mathbf{\Lambda}$ is a diagonal matrix of the eigenvalues sorted in the descending order,

$$\mathbf{\Lambda} = \mathrm{diag}\left[\lambda_0, \cdots, \lambda_{D-1}, \cdots, \lambda_{N-2}\right], \tag{1.103}$$

and $\mathbf{E}$ is a matrix of the corresponding eigenvectors,

$$\mathbf{E} = \left[\mathbf{e}_0, \cdots, \mathbf{e}_{D-1}, \cdots, \mathbf{e}_{N-2}\right]. \tag{1.104}$$

Here, we assume that $\Sigma_{\mathbf{N}}$ is an identity matrix. Then, we select the eigenvectors with the D largest eigenvalues, $\mathbf{E}_V = [\mathbf{e}_0, \cdots, \mathbf{e}_{D-1}]$. In the similar manner, we define the subspace for the ambient noise as $\mathbf{E}_N = [\mathbf{e}_D, \cdots, \mathbf{e}_{N-2}]$.

The ideal properties of the eigenvectors and eigenvalues can be summarized as follows.

- The subspace spanned by the eigenvectors is equal to that of $\mathbf{A}$, i.e., $\Re\{\mathbf{E}_V\} = \Re\{\mathbf{A}\}$.
- The power of the D spatially correlated signals is associated with the D largest eigenvalues.
- The power of $\mathbf{N}(k)$ is equally spread over all the eigenvalues and $N - D - 1$ smallest eigenvalues are all equal to σ_N^2, i.e., the noise floor.
- $\mathbb{R}\{\mathbf{E}_N\}$ is the orthogonal complement of $\mathbb{R}\{\mathbf{E}_V\}$, i.e., $\mathbb{R}\{\mathbf{E}_N\} = \mathbb{R}\{\mathbf{E}_V\}^{\perp}$.

In order to cluster the eigenvectors for the ambient noise, we have to determine the number of the dominant eigenvalues D. Figure 1.13 illustrates actual eigenvalues sorted in descending order over frequencies. In order to generate the plots of the figures in this section, we computed the eigenvalues from the outputs of the blocking matrix on the real data described in Kumatani et al. (2011).

As shown in Figure 1.13, it is relatively easy to determine the number of the dominant modes, D, especially in the case that the number of the microphones is much larger than

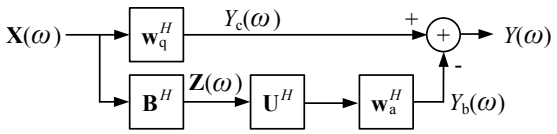


Figure 1.15 Maximum kurtosis beamformer with the subspace filter.

the number of the spatially correlated signals. We determine D based on the threshold of the contribution ratio, $\lambda_i / \sum_{j=0}^{N-2} \lambda_j$. Figure 1.14 shows averages of numbers of the contribution ratios exceeding thresholds, 10^{-2}, 10^{-3} and 10^{-4}, at each frequency. Figure 1.14 indicates how many dominant modes are used for subspace filtering when we ignore the eigenvectors associated with the lower contribution ratio than the threshold. It is clear from Figure 1.14 that the lower the threshold for the contribution ratio is set, the more eigenvectors are used. Generally, the lower threshold leads to accurate representation of the spatially correlated signals at the expense of computational efficiency.

In the optimization of the active weight vector, we estimate each component corresponding to the ambient noise signal separately. Accordingly, we use the sum of the eigenvectors for the ambient noise space as $\tilde{\mathbf{e}}_n = \sum_{d=D}^{N-2} \mathbf{e}_d$. Our subspace filter can be now written as

$$\mathbf{U} = [\mathbf{e}_0, \cdots, \mathbf{e}_{D-1}, \tilde{\mathbf{e}}_n,]\,. \tag{1.105}$$

Note that we assume the covariance matrix in (1.102) can be approximated as

$$\Sigma_{\mathbf{Z}} \approx \sum_{d=0}^{D-1} \lambda_d \mathbf{e}_d \mathbf{e}_d^H + \bar{\sigma}_N^2 \tilde{\mathbf{e}}_n \tilde{\mathbf{e}}_n^H, \tag{1.106}$$

where

$$\bar{\sigma}_N^2 = \frac{1}{N-D-1} \sum_{d=D}^{N-2} \lambda_d$$

With the outputs of the subspace filter, we estimate the active weight vector providing the maximum kurtosis value. If the output of the subspace filter is a noise signal, the corresponding component of the active weight vector should be adjusted so as to subtract the noise component from the output of the quiescent vector. If it is an echo of the desired signal, the active weight vector could shift the phase and add the component to the desired signal in order to strengthen it. These operations would be easier by separating the echo from the ambient noise component with the subspace filter.

Block-Wise Adaptation of Maximum Kurtosis Beamforming with Subspace Filtering

Figure 1.15 shows configuration of the MK beamformer with the subspace filter. The beamformer's output can be expressed as

$$Y(t) = \left[\mathbf{w}_{\mathrm{q}}(t) - \mathbf{B}(t)\mathbf{U}(t)\mathbf{w}_{\mathrm{a}}(t)\right]^H \mathbf{X}(t). \tag{1.107}$$

The active weight vector is adjusted so as to achieve the maximum kurtosis of the beamformer's outputs. The difference between (1.74) and (1.107) is the subspace filter between the blocking matrix and active weight vector. The subspace filter can decompose the output vector into the spatially correlated signal and ambient noise components. Therefore, we only need to estimate the phase shifts of the active weight vector on the constrained subspace (Hyvärinen and Oja 2000, §7.4). Moreover, the solution of the general eigenvector decomposition is less dependent of the initial values than that of the gradient algorithm for multi-dimensional maximization or minimization.

Based on equation (1.107), the kurtosis of the outputs is computed from a block of input subband samples at each block instead of using the entire utterance data. We incrementally update the dominant modes and active weight vector at each block b consisting of L_b samples here. Accordingly, the kurtosis at block b can be expressed as

$$J(Y(b)) = \left(\frac{1}{L_b} \sum_{t=(b-1)L_b}^{bL_b-1} |Y(t)|^4 \right) - \beta \left(\frac{1}{L_b} \sum_{k=(b-1)L_b}^{bL_b-1} |Y(t)|^2 \right)^2 . \tag{1.108}$$

In order to inhibit the formation of excessively large sidelobes, a regularization term is applied to the cost function as follows:

$$\mathcal{J}(Y(b);\alpha) = J(Y(b)) - \alpha \|\mathbf{w}_{\mathrm{a}}(b)\|^2 \tag{1.109}$$

In addition to the regularization term, a unity constraint is imposed on a norm of the active weight vector so as to prevent it from exceeding that of the quiescent vector.

Then, the active weight vector is adjusted so as to maximize the sum of the kurtosis and regularization terms (1.109) under the norm constraint at each block. Again, we have to resort to the gradient-based numerical optimization algorithm. Upon substituting (1.107) and (1.108) into (1.109) and taking the partial derivative with respect to the active weight vector, we obtain

$$\begin{aligned} \frac{\partial \mathcal{J}(Y(b);\alpha)}{\partial \mathbf{w}_{\mathrm{a}}(b)^*} &= -\frac{2}{L_b} \left(\sum_{t=(b-1)L_b}^{bL_b-1} |Y(t)|^2 \mathbf{U}^H(b) \mathbf{B}^H(t) \mathbf{X}(t) Y^*(t) \right) \\ &+ \frac{2\beta}{L_b^2} \left(\sum_{t=(b-1)L_b}^{bL_b-1} |Y(t)|^2 \right) \left(\sum_{t=(b-1)L_b}^{bL_b-1} \mathbf{U}^H(b) \mathbf{B}^H(t) \mathbf{X}(t) Y^*(t) \right) - \alpha \mathbf{w}_{\mathrm{a}}(b). \end{aligned} \tag{1.110}$$

The gradient (1.110) is iteratively calculated with a block of subband samples until the kurtosis value of the beamformer's outputs converges. For the gradient algorithm, the active weight vectors are initialized with the estimates at the previous block. The active weight vector of the first block is initialized with $\mathbf{w}_{\mathrm{a}} = [0, \cdots, 0, 1]^T$ because the last component corresponds to the ambient noise which should be subtracted from the output of the quiescent vector .

The beamforming algorithm can be summarized as follows:

1. Initialize the active weight vector with $\mathbf{w}_{\mathrm{a}}(0) = [0, \cdots, 1]^T$.
2. Given estimates of time delays, calculate the quiescent vector and blocking matrix.

3. For each block of input subband samples, recursively update the covariance matrix as $\Sigma_{\mathbf{Z}}(b) = \mu\Sigma_{\mathbf{Z}}(b-1) + (1-\mu)\Sigma_{\mathbf{Z}}(b)$ where μ is the forgetting factor, calculate the dominant modes $\mathbf{U}^H(b)$ and estimate the active weight vector $\mathbf{w}_{\mathrm{a}}(b)$ based on the gradient information computed with (1.110) subject to the norm constraint until the kurtosis value of the beamformer's outputs converges.
4. Initialize the active weight vector obtained in step 3 for the next block and go to the step 2.

This block-wise method is able to track a non-stationary sound source, and provides a more accurate gradient estimate than *sample-by-sample* gradient estimation algorithms.

1.3.10 Speech Recognition Experiments

The results of the distant speech recognition (DSR) experiments reported in this section were obtained on speech material from the Multi-Channel Wall Street Journal Audio Visual Corpus (MC-WSJ-AV) recorded by the Augmented Multi-party Interaction (AMI) project; see Lincoln et al. (2005) for details of the data collection apparatus. The size of the recording room was $650 \times 490 \times 325$ cm and the reverberation time T_{60} was approximately 380 ms. In addition to reverberation, some recordings include significant amounts of background noise produced by computer fans and air conditioning. The far-field speech data was recorded with two circular, equi-spaced eight-channel microphone arrays with the diameter of 20cm. Additionally, each speaker was equipped with a close talking microphone (CTM) to provide a reference signal for speech recognition. The sampling rate of the recordings was 16 kHz. For the experiments, we used a portion of data from the *single speaker stationary* scenario where a speaker was asked to read sentences from six positions, four seated around the table, one standing at the white board and one standing at a presentation screen. The test data set for the experiments contains recordings of 10 speakers where each speaker reads approximately 40 sentences taken from the 5,000 word vocabulary WSJ task. This provided a total of 352 utterances for a total 39.2 minutes of speech.

Prior to beamforming, we first estimated the speaker's position with a source tracking system (Gehrig et al. 2006). Based on the average speaker position estimated for each utterance, active weight vectors $\mathbf{w}_a$ were estimated for the source. After beamforming, we perform *Zelinski post-filtering* (Marro et al. 1998) which uses the auto- and cross-power spectrums of the input signals to estimate the target signal and noise power spectrums under the assumption of zero cross-correlation between noise on different sensors. The parameters of the GG pdf were trained with 43.9 minutes of speech data recorded with the CTM in the SSC2 development set. The training data set for the GG pdf contains recordings of 5 speakers.

We performed four decoding passes on the waveforms obtained with various beamforming algorithms including ones described in prior sections. The details of our ASR system used in the experiments are given in Kumatani et al. (2008a). Each pass of decoding used a different acoustic model or speaker adaptation scheme. The speaker adaptation parameters were estimated using the word lattices generated during the prior pass, as in Uebel and Woodland (2001). A description of the four decoding passes follows:

1. Decode with the unadapted, conventional ML acoustic model and bigram language model (LM).

Table 1.1 Word error rates for each beamforming algorithm after every decoding pass.

Beamforming (BF) Algorithm	Pass (%WER)			
	1	2	3	4
Single array channel (SAC)	87.0	57.1	32.8	28.0
Delay-and-sum (D&S) BF	79.0	38.1	20.2	16.5
Superdirective (SD) BF	71.4	31.9	16.6	14.1
Minimum variance distortionless response (MVDR) BF	78.6	35.4	18.8	14.8
Generalized eigenvector (GEV) BF	78.7	35.5	18.6	14.5
Maximum kurtosis (MK) BF	75.7	32.8	17.3	13.7
Maximum negentropy (MN) BF	75.1	32.7	16.5	13.2
SD MN BF	75.3	30.9	15.5	12.2
Close talking microphone (CTM)	52.9	21.5	9.8	6.7

2. Estimate vocal tract length normalization (VTLN) (Wölfel and McDonough 2009, §9) parameters and *constrained maximum likelihood linear regression* (CMLLR) parameters for each speaker as discussed in Section **??**, then redecode with the conventional ML acoustic model and bigram LM.
3. Estimate VTLN, CMLLR, and *maximum likelihood linear regression* (MLLR) parameters for each speaker as discussed in Section **??**, then redecode with the conventional model and bigram LM.
4. Estimate VTLN, CMLLR, MLLR parameters for each speaker, then redecode with the ML-SAT model (Wölfel and McDonough 2009, §8.1) and bigram LM.

Table 1.1 shows the word error rates (WERs) for every beamforming algorithm. As references, WERs in recognition experiments on speech data recorded with a *single array channel* (SAC) and CTM are also presented in the table. It is clear from these results that the maximum kurtosis beamforming (MK BF) and maximum negentropy beamforming (MN BF) methods can provide better recognition performance than the SOS-based beamformers, such as a superdirective beamformer (SD BF) (Wölfel and McDonough 2009, §13.3.4), the MVDR beamformer (MVDR BF) described in Section 1.3.1, and the generalized eigenvector beamformer (GEV BF) (Warsitz et al. 2008). This is because the HOS-based beamformers can use echoes to enhance the desired signal, as mentioned previously. Adaptation of the HOS-based beamformers was performed while the desired source was active. This fact implies that the maximum kurtosis and negentropy beamformers do not suffer the signal cancellation. It is also clear from Table 1.1 that every adaptive beamformer achieved better recognition performance than the delay-and-sum beamformer (D&S BF).

The SOS-based and HOS-based beamformers can be profitably combined because maximum kurtosis and negentropy beamformers employ different criteria for estimation of the active weight vector. For example, the superdirective beamformer's weight can be used as the quiescent weight vector in GSC configuration Kumatani et al. (2010a). We observe from Table 1.1 that the maximum negentropy beamformer with superdirective beamformer (SD MN BF) provided the best recognition performance in this task.

1.4 Spherical Microphone Arrays

In this section we discuss the fundamentals of spherical arrays. This includes the acoustic phenomena that occur when a plan wave scatters from the surface of a rigid sphere. We also develop the concept of expanding a function defined on the surface of a sphere in spherical harmonics; such series expansions will play a key role in our development of beamforming algorithms for spherical arrays. This material is intended to provide the theoretical underpinning to the emprical studies presented in the next section.

Meyer and Elko (2002) and Abhayapala and Ward (2002) were among the first to propose the use of spherical microphone arrays for beamforming. The state-of-the-art theory of beamforming with spherical microphone arrays explicitly takes into account two phenomena of sound propagation, namely, *diffraction* and *scattering*; see Kutruff (2009, §2) and Williams (1999, §6.10). While these phenomena are certainly present in all acoustic array processing applications, no particular attempt is typically made to incorporate them into conventional beamforming algorithms; rather, they are simply assumed to contribute to the room impulse response. The development in this section is based loosely on Meyer and Elko (2004, §2) with interspersed elements from Teutsch (2007).

To begin our discussion, let us express a plane wave impinging with a polar angle of θ on an array of microphones as

$$G_{\text{pw}}(kr, \theta, t) \triangleq e^{i(\omega t + kr\cos\theta)}, \tag{1.111}$$

where $k \triangleq 2\pi/\lambda$ is the wavenumber as before, and r is the range at which the wave is observed. The definition (1.111) can be rewritten as

$$G_{\text{pw}}(kr, \theta, t) = \sum_{n=0}^{\infty} i^n \, (2n+1) \, j_n(kr) \, P_n(\cos\theta) \, e^{i\omega t}, \tag{1.112}$$

where j_n and P_n are respectively the *spherical Bessel function* of the first kind (Olver and Maximon 2010, §10.47) and the *Legendre polynomial*, both of order n (Williams 1999, §6.10.1). A similar expansion of spherical waves—such as would be required for near-field analysis—is provided by Williams (1999, §6.7.1). If the plane wave encounters a rigid sphere with a radius of a it is *scattered* (Williams 1999, §6.10.3) to produce a wave with the pressure profile

$$G_{\text{s}}(kr, ka, \theta, t) = -\sum_{n=0}^{\infty} i^n \, (2n+1) \, \frac{j_n'(ka)}{h_n'(ka)} \, h_n(kr) \, P_n(\cos\theta) \, e^{i\omega t}, \tag{1.113}$$

where $h_n = h_n^{(1)}$ denotes the *Hankel function* (Olver and Maximon 2010, §10.47) of the first kind[2] while the prime indicates the derivative of a function with respect to its argument. Combining (1.112) and (1.113) yields the total sound pressure field (Williams 1999, §6.10.3)

$$G(kr, ka, \theta) = \sum_{n=0}^{\infty} i^n (2n+1) \, b_n(ka, kr) \, P_n(\cos\theta), \tag{1.114}$$

[2]Note that Meyer and Elko (2004) incorrectly used the Hankel function of the second kind in (1.113) and (1.115).

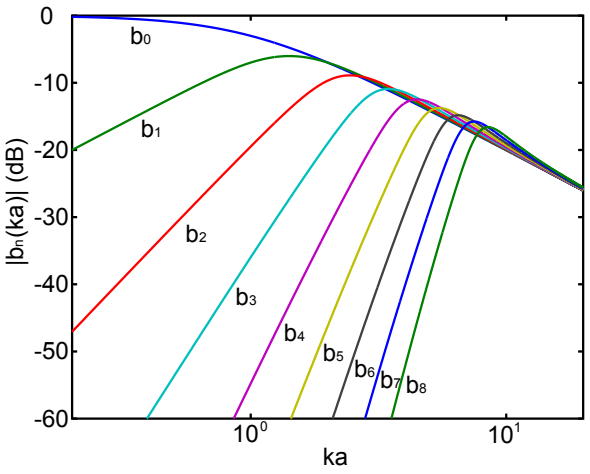


Figure 1.16 Magnitudes of the modal coefficients $b_n(ka, ka)$ for $n = 0, 1, \ldots, 8$, where a is the radius of the scattering sphere and k is the wavenumber.

where the nth *modal coefficient* is defined as

$$b_n(ka, kr) \triangleq j_n(kr) - \frac{j_n'(ka)}{h_n'(ka)} h_n(kr). \tag{1.115}$$

In principle, ka and kr need not be equivalent, but in practice they are; i.e., the sensors of the array are located on the surface of the scattering sphere. Hence, in the sequel we will uniformly replace kr with ka. Note that the time dependence of (1.114) through the term $e^{i\omega t}$ has been suppressed for convenience. Plots of $|b_n(ka, ka)|$ for $n = 0, \ldots, 8$ are shown in Figure 1.16.

Let us now define the *spherical harmonic* of order n and degree m as (Driscoll and Dennis M. Healy 1994)

$$Y_n^m(\Omega) \triangleq \sqrt{\frac{(2n+1)}{4\pi} \frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta)\, e^{im\phi}, \tag{1.116}$$

where $\Omega \triangleq (\theta, \phi)$ and P_n^m is the *associated Legendre function* of order n and degree m (Dunster 2010, §14.3). The spherical harmonics fulfill the same role in the decomposition of square-integrable functions defined on the surface of a sphere as that played by the complex exponential $e^{i\omega nt}$ for Fourier analysis of periodic functions defined on the real line. Let γ represent the angle between the points $\Omega \triangleq (\theta, \phi)$ and $\Omega_s \triangleq (\theta_s, \phi_s)$ lying on a sphere, such that

$$\cos\gamma = \cos\theta_s \cos\theta + \sin\theta_s \sin\theta \cos(\phi_s - \phi). \tag{1.117}$$

Then we can express the *addition theorem for spherical harmonics* (Arfken and Weber 2005, §12.8) as

$$P_n(\cos\gamma) = \frac{4\pi}{2n+1} \sum_{m=-n}^{n} Y_n^m(\Omega_s) \bar{Y}_n^m(\Omega), \tag{1.118}$$

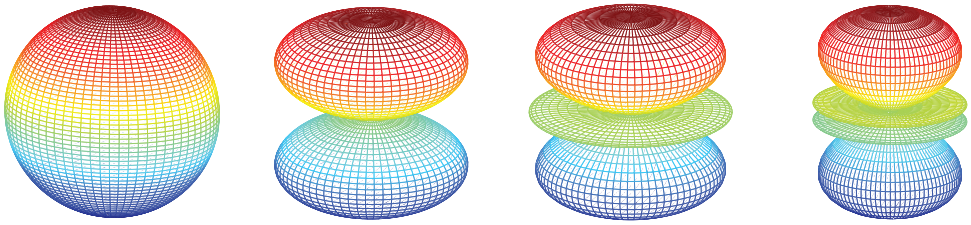


Figure 1.17 The spherical harmonics Y_0, Y_1, Y_2 and Y_3.

where $\bar{Y}$ denotes the complex conjugate of Y. Upon substituting (1.118) into (1.114), we find

$$G(\Omega_s, ka, \Omega) = 4\pi \sum_{n=0}^{\infty} i^n \, b_n(ka) \sum_{m=-n}^{n} Y_n^m(\Omega_s) \bar{Y}_n^m(\Omega). \tag{1.119}$$

The spherical harmonics $Y_0 \triangleq Y_0^0$, $Y_1 \triangleq Y_1^0$, $Y_2 \triangleq Y_2^0$ and $Y_3 \triangleq Y_3^0$ are shown in Figure 1.17. The spherical harmonics possess the all important property of *orthonormality*, which implies

$$\begin{aligned} \delta_{n,n'} \, \delta_{m,m'} &= \int_{\Omega} Y_n^m(\Omega) \, \bar{Y}_{n'}^{m'}(\Omega) \, d\Omega \\ &= \int_0^{2\pi} \int_0^{\pi} Y_n^m(\Omega) \, \bar{Y}_{n'}^{m'}(\Omega) \sin\theta d\theta d\phi, \end{aligned} \tag{1.120}$$

where Ω indicates the surface of the sphere of integration and the *Kronecker delta* is defined as

$$\delta_{n,m} \triangleq \begin{cases} 1, & n, m = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{1.121}$$

The plot of Y_0 in Figure 1.17 and magnitudes of the modal coefficients in Figure 1.16 make it clear why the spherical array—like its linear, conventional counterpart—suffers from poor directivity at low frequencies. For $ka = 0.2$, which corresponds to 220 Hz for $a = 5$ cm, only Y_0 is truly useful for beamforming, as the levels of all other modes are 20 dB or more below that of Y_0; amplifying the other modes sufficiently to use them in beamforming would introduce a great deal of self-noise from the array components into the final signal. But Y_0 is completely isotropic; i.e., it has no directional characteristics whatsoever, and hence provides no improvement in directivity over a single microphone.

The implication of (1.120) is that the individual terms of the series expansion are orthonormal. Hence, any sound field $V(ka, \Omega_s)$, which is square-integrable over a sphere with radius a, admits the modal decomposition (Teutsch 2007, §A.3),

$$V(ka, \Omega_s) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} V_n^m(ka) \, Y_n^m(\Omega_s), \tag{1.122}$$

when observed at (a, Ω_s), where

$$V_n^m(ka) \triangleq \int_{\Omega_s} V(ka, \Omega_s) \, \bar{Y}_n^m(\Omega_s) \, d\Omega_s \tag{1.123}$$

is the (n, m)th coefficient of the decomposition. Equation (1.122) is readily verified by substituting (1.122) into (1.123) and applying the orthonormality property (1.120). The coefficients $V_n^m(ka)$ represent a transform domain much like the Fourier coefficients of a periodic function.

Specializing the above by substituting $V(\Omega_s) = G(\Omega_s, ka, \Omega)$ from (1.119) into (1.123) yields

$$G_n^m(\Omega, ka) = \int_{\Omega_s} G(\Omega_s, ka, \Omega)\, \bar{Y}_n^m(\Omega_s)\, d\Omega_s \tag{1.124}$$

$$= 4\pi \sum_{n'=0}^{\infty} i^{n'} b_{n'}(ka) \sum_{m'=-n'}^{n'} \bar{Y}_{n'}^{m'}(\Omega) \int_{\Omega_s} Y_{n'}^{m'}(\Omega_s) \bar{Y}_n^m(\Omega_s)\, d\Omega_s$$

$$= 4\pi\, i^n b_n(ka)\, \bar{Y}_n^m(\Omega), \tag{1.125}$$

where the latter equality follows directly from (1.120).[3] These results will shortly prove useful in deriving the modal analog of the array manifold vector defined in Section 1.2.

Equation (1.122) can be interpreted as the decomposition of an arbitrary square-integrable sound field into an infinite series of spherical harmonics or *modes*. Equation (1.124) is then a specialization for a plane wave. In the next section, we will consider how such decompositions can be used for beamforming.

For the case of a discrete array of microphones as opposed to a continuous, pressure-sensitive surface, it is necessary to reformulate (1.120) as (Li et al. 2004)

$$\frac{4\pi}{S} \sum_{s=0}^{S-1} Y_n^m(\Omega_s)\, \bar{Y}_{n'}^{m'}(\Omega_s) = \delta_{n,n'} \delta_{m,m'}, \tag{1.126}$$

where S is the number of sensors, each of which is located at some (Ω_s) for $s = 0, 1, \ldots, S - 1$. Similarly, we can define a discrete version of (1.123) as

$$V_n^m(ka) \triangleq \frac{4\pi}{S} \sum_{s=0}^{S-1} V(ka, \Omega_s)\, \bar{Y}_n^m(\Omega_s), \tag{1.127}$$

and of the modal decomposition of the plane wave (1.125) as

$$G_n^m(\Omega, ka) = \frac{4\pi}{S} \sum_{s=0}^{S-1} G(\Omega_s, ka, \Omega)\, \bar{Y}_n^m(\Omega_s). \tag{1.128}$$

Note that the presence of the leading coefficient of $4\pi/S$ in (1.126–1.128) ensures that (1.123) and (1.127) are equivalent for $V_n^m(ka) \equiv 1$. Both instruments are available commercially and have attracted a great deal of interest within the acoustic beamforming research community.

Equations (1.127) and (1.128) are of fundamental importance in that they define the *modal decomposition* that is typically performed prior to beamforming with a spherical array. One of the primary challenges in designing realizeable and effective spherical arrays is choosing

[3] Teutsch (2007, §5.1.2) incorrectly reports the leading coefficient of (1.125) as $\sqrt{4\pi}$.

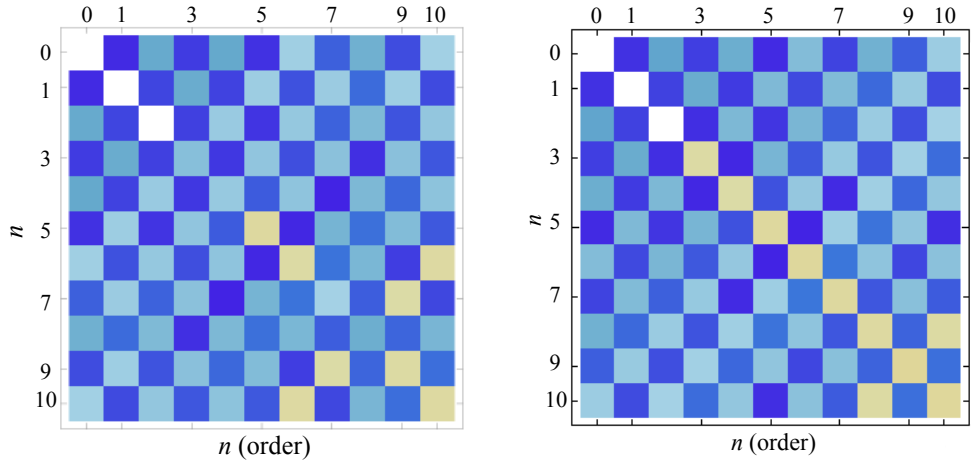


Figure 1.18 Deviation of the spherical harmonics from orthonormality for a) $S = 24$ and b) $S = 32$ elements.

the set of sensor locations $\{(\Omega_s)\}$ such that (1.126) is satisfied as nearly as possible; see Li and Duraiswami (2005); Li et al. (2004). Li and Duraiswami (2007) discuss the use of variable *quadrature weights* (Fliege and Maier 1999) to minimize the orthonormality error between (1.120) and (1.126). The early work by Meyer and Elko (2004) reported two sensor placement schemes for arrays of $S = 24$ and 32 elements; the deviation of the modes of these discrete arrays from orthonormality are illustrated in Figure 1.18, wherein lighter colored squares indicate a higher sensitivity reported in a decibel scale. From Figure 1.18 a) it is clear that the 24-element array maintains orthogonality only through mode $n = 2$, while Figure 1.18 b) indicates that the 32-element array maintains orthogonality through $n = 8$. Typically, however, the order is truncated such that $(N + 1)^2 \leq S$ to avoid spatial aliasing, implying that a 32-element array can support a maximum order of $N = 4$.

Shown in Figure 1.19 are a spherical, 32-element *Eigenmike*® manufactured by *mh acoustics* and a 64-element, spherical array with five integrated video cameras manufactured by *VisiSonics Corporation*.

a) b)

Figure 1.19 a) A 32-channel Eigenmike® spherical array; photo reproduced courtesy of *mh acoustics, Inc.* b) A 64-channel spherical array with five integrated video cameras; photo reproduced courtesy of *VisiSonics Corporation.*

Stacking the modal components (1.125) together provides the *modal array manifold vector*,

$$\mathbf{v}(\Omega, ka) \triangleq \begin{bmatrix} G_0^0(\Omega, ka) \\ G_1^{-1}(\Omega, ka) \\ G_1^0(\Omega, ka) \\ G_1^1(\Omega, ka) \\ G_2^{-2}(\Omega, ka) \\ G_2^{-1}(\Omega, ka) \\ G_2^0(\Omega, ka) \\ \vdots \\ G_N^{-N}(\Omega, ka) \\ \vdots \\ G_N^N(\Omega, ka) \end{bmatrix}, \tag{1.129}$$

where Ω denotes the direction of arrival of a plane wave. The modal array manifold vector is so dubbed because it fulfills precisely the same role as the array manifold vector defined in (1.29); i.e., it describes the excitation of the (n, m)th mode—as opposed to the nth *microphone*—of the spherical array by a plane wave arriving from the direction Ω; clearly this interaction is more complicated than the simple phase shift seen in the conventional array.

Evaluating any individual component $G_n^m(\Omega, ka)$ in $\mathbf{v}(\Omega, ka)$ requires (1.125) as well as the identity

$$Y_n^{-m}(\Omega) \equiv (-1)^m \bar{Y}_n^m(\Omega);$$

the latter follows directly from the definition (1.116) and the identity (Dunster 2010, §14.9)

$$P_n^{-m}(x) \equiv (-1)^m \frac{(n-m)!}{(n+m)!} P_n^m(x).$$

With these relations in mind, we can trivially evaluate $G_3^{-2}(\Omega, ka)$, for example, as

$$G_3^{-2}(\Omega, ka) = 4\pi\, i^3 b_3(ka) \cdot (-1)^2\, Y_3^2(\Omega) = -4\pi\, i\, b_3(ka)\, Y_3^2(\Omega).$$

To close this section, we note that while spherical arrays possess several attractive characteristics, they are no panacea for the many maladies of distant speech recognition; they too suffer from poor directivity at low frequencies and spatial aliasing at high frequencies just like conventional arrays. Establishing the suitability of spherical arrays for distant speech recognition will require both more detailed analysis and empirical studies; we turn our attention to both of these tasks in the coming sections.

1.5 Spherical Adaptive Algorithms

In this case we investigate the well-known *minimum variance distortionless response* (MVDR) beamformer for spherical arrays. The solution for the MVDR beamforming weights with diagonal loading is given by (1.41). 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 Meyer and Elko (2002, 2004).

For use in the GSC discussed in Section 1.3.5, we can set

$$\mathbf{w}_{\mathrm{q}}(\Omega, ka) = \frac{1}{C}\mathbf{v}(\Omega, ka), \tag{1.130}$$

where C is a normalization constant that ensures satisfaction of the distortionless constraint,

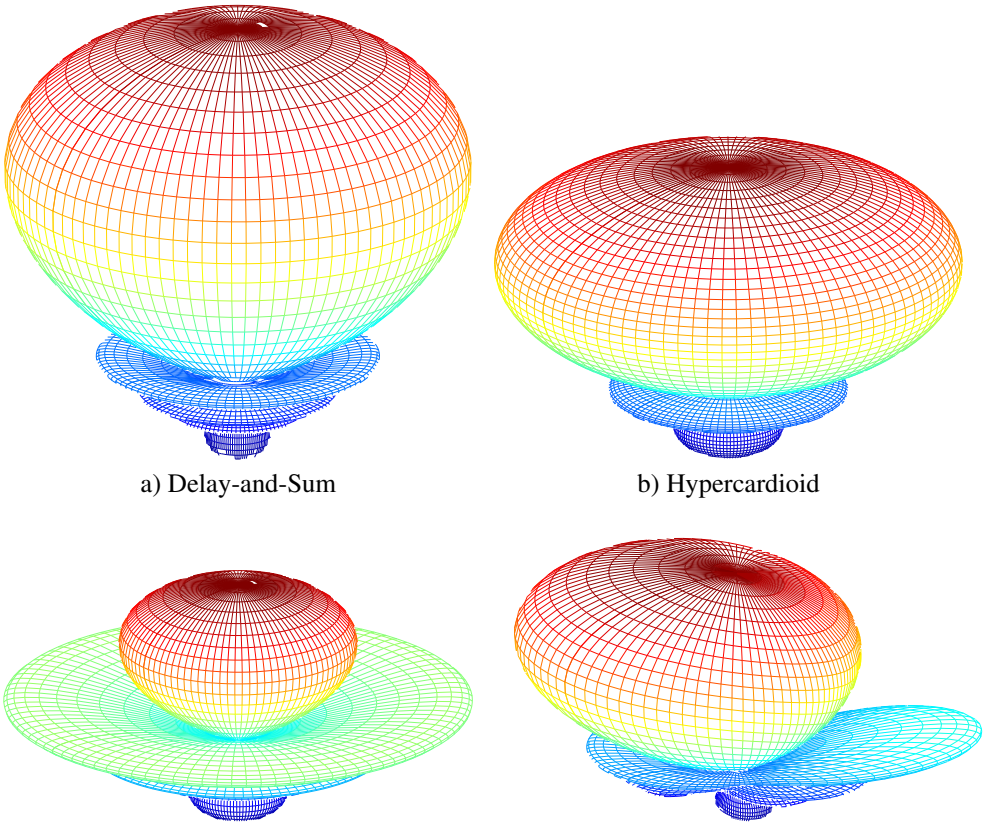
$$\mathbf{w}_{\mathrm{q}}^H(\Omega_0, ka)\, \mathbf{v}(\Omega_0, ka) = 1, \tag{1.131}$$

for the look direction Ω_0. The blocking matrix can then be derived in the normal way from $\mathbf{w}_{\mathrm{q}}^H(\theta, \phi, ka)$.

With formulations of the relevant array manifold vector (1.129), we can immediately write the solution (1.30) for the *delay-and-sum* beamformer. Another popular fixed design for spherical array processing is the hypercardioid (Meyer and Elko 2004). The beampatterns obtained with the delay-and-sum and hypercardioid designs are shown in Figure 1.20 a) and b) respectively. The MVDR design both with and without a radial symmetry constraint are shown in Figure1.20 c) and d), respectively.

Let us now consider the case where the look direction is $(\theta, \phi) = (0, 0)$ and there is a single strong interference signal impinging on the array from $(\theta_{\mathrm{I}}, \phi_{\mathrm{I}}) = (\pi/6, 0)$ with a magnitude of $\sigma_{\mathrm{I}}^2 = 10^{-1}$. In this case, the covariance matrix of the array input is

$$\Sigma_{\mathbf{X}}(ka) = \mathbf{v}(\Omega, ka)\, \mathbf{v}^H(\Omega, ka) + \sigma_{\mathrm{I}}^2\, \mathbf{v}(\theta_{\mathrm{I}}, \phi_{\mathrm{I}}, ka)\, \mathbf{v}^H(\theta_{\mathrm{I}}, \phi_{\mathrm{I}}, ka). \tag{1.132}$$



a) Delay-and-Sum b) Hypercardioid

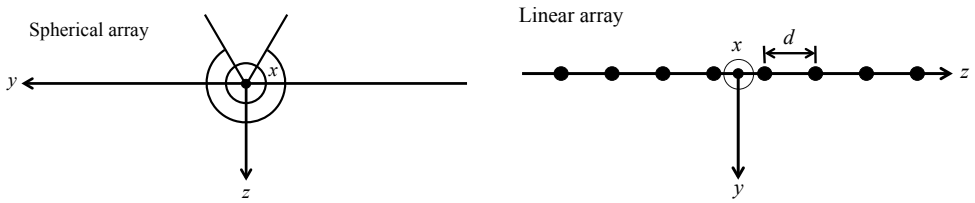
c) Radially Symmetric MVDR d) Asymmetric MVDR

Figure 1.20 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.

1.6 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 1.3.4, 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 1.21 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

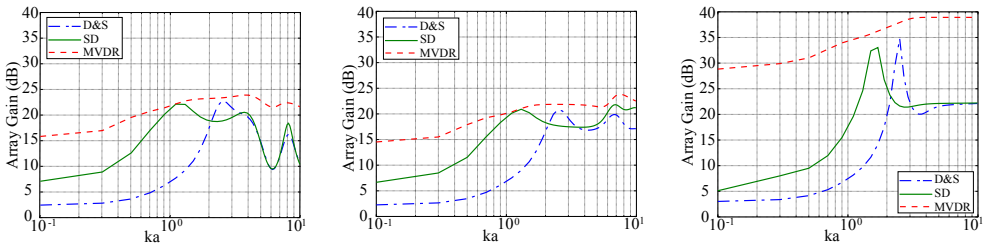


a) Spherical 32-channel array. b) Conventional, 64-channel, linear array.

Figure 1.21 Orientation of the a) Mark IV linear array and b) Eigenmike® spherical array.

Table 1.2 Acoustic environment for comparing the Mark IV linear array with the Eigenmike® spherical microphone 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

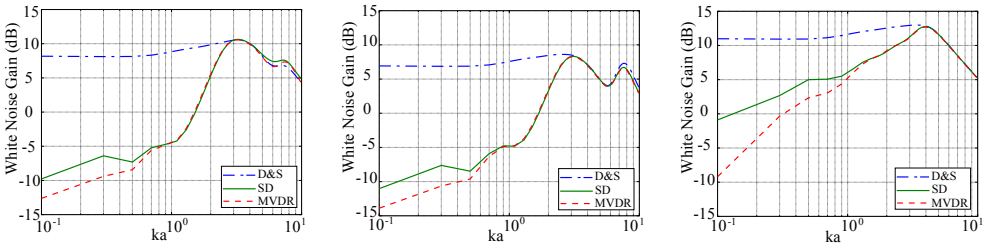


a) $S = 24$ b) $S = 32$ c) Ideal array

Figure 1.22 Array gain as a function of ka for a) $S = 24$, b) $S = 32$, and c) the Ideal array.

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 1.2. The specific arrays we chose to simulate were the Mark IV linear array and the Eigenmike® spherical array.

In Figure 1.22 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 1.3.1, 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



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

Figure 1.23 White noise gain as a function of ka for a) $S = 24$, b) $S = 32$, and c) the Ideal array.

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 1.23 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 $ka \geq 30$.

As described in Section 1.2, the beampattern is the sensitivity of the array to a plane wave arriving from some direction Ω. By weighting each spherical mode (1.125) 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 (1.70) is given by

$$P(\Omega) \triangleq |B(\Omega, ka)|^2 \tag{1.133}$$

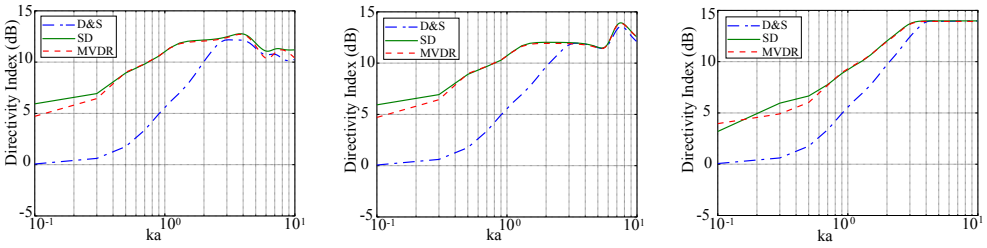
$$= 16\pi^2 \sum_{n,n'=0}^{\infty} i^n \, \bar{i}^{n'} \, b_n(ka) \, \bar{b}_{n'}(ka) \sum_{m,m'=-n,-n'}^{n,n'} \bar{w}_n^m \, w_{n'}^{m'} \, \bar{Y}_n^m(\Omega) \, Y_{n'}^{m'}(\Omega).$$

Substituting (1.133) into (1.72) and applying (1.120) then provides (Yan et al. 2011)

$$\mathrm{DI}_{\mathrm{ideal}}(ka, \mathbf{w}) = -10 \log_{10} \left\{ 4\pi \sum_{n=0}^{N} |b_n(ka)|^2 \sum_{m=-n}^{m} |w_n^m|^2 \right\}. \tag{1.134}$$

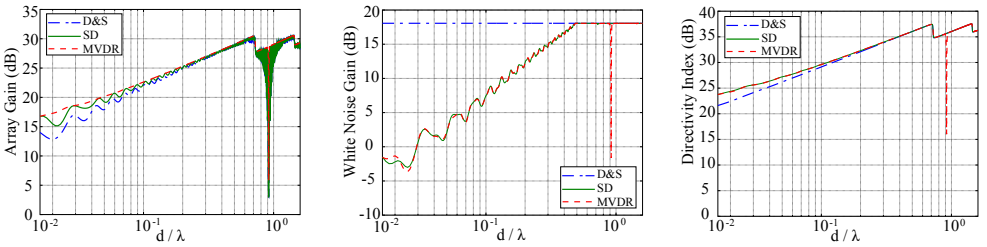
The directivity index as a function of ka for both ideal and discrete arrays is plotted in Figure 1.24. These figures reveal that—as anticipated by the theory of Section 1.3.1—the superdirective (SD) beamformer provides the highest directivity save in the very low frequency region where the sensor covariance matrix (1.42) 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 1.25, where each metric is shown as a function of d/λ, the ratio of intersensor spacing



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

Figure 1.24 Directivity index as a function of ka for a) $S = 24$, b) $S = 32$, and c) the Ideal array.



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

Figure 1.25 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.

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 as discussed in Section 1.2.

1.7 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, 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 1.26 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,

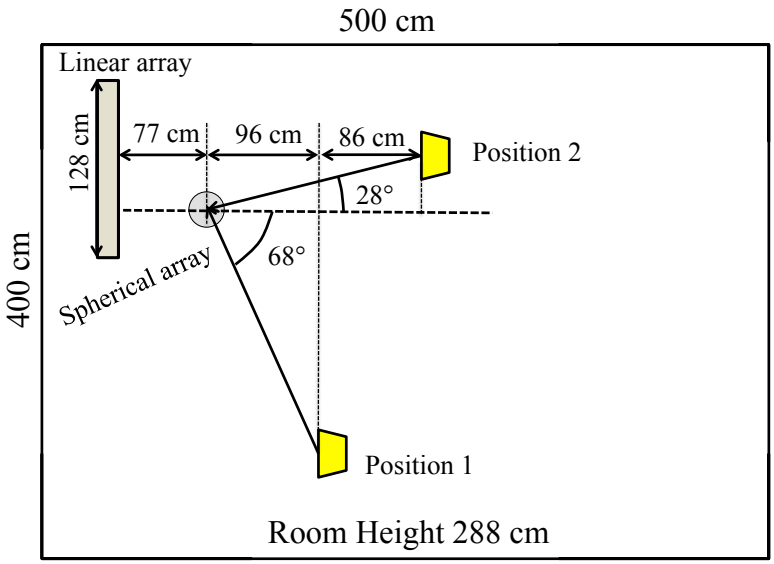


Figure 1.26 The layout of the recording room.

Table 1.3 WERs for each beamforming algorithm in the case that the incident angle to the array is 28^o.

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
SD BF with spherical array	43.9	14.2	12.1	10.5
Spherical D&S BF	48.2	17.1	13.9	12.6
Spherical SD BF	43.0	14.6	11.5	9.9
CTM	16.7	7.5	6.4	5.4

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 recognizer described in Section 1.3.10. Tables 1.3 and 1.4 show word error rates (WERs) for each beamforming algorithm for the cases wherein the incident

Table 1.4 WERs for each beamforming algorithm in the case that the incident angle to the array is 68^o.

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	60.0	26.1	18.1	16.2
Spherical SD BF	45.6	15.3	12.6	10.7
CTM	16.7	7.5	6.4	5.4

angles 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 they it is able to maintain the same beam pattern regardless of the incident angle. In these experiments, spherical D&S beamforming (Spherical D&S BF) could not improve the recognition performance significantly because of its poor directivity.

1.8 Conclusions and Further Reading

In this contribution, we have examined the application of spherical microphone arrays to beamforming and compared this with the use of conventional arrays. As we have explained, the primary difference between the conventional and spherical array processing literature is that the latter makes an explicit attempt to account for and model two phenomena that are present in all acoustic array processing applications, namely, diffraction, which is the tendency of sound to bend around fixed obstacles, and scattering, which is the tendency of sound to be dispersed through reflection from non-planar surfaces. By modeling both phenomena, spherical array processing typically begins by decomposing the sound field into a number of orthogonal components, which are subsequently weighted and combined, much like the outputs of single microphones in conventional array processing. The advantage of the spherical configuration is that it is spatially isotropic, implying that the look direction can be set to nearly any spherical coordinate with equal ease.

We have found that both conventional and spherical arrays suffer from two primary problems:

1. Poor low-frequency directivity, which is a result of the finite physical aperture that a realizeable array must necessarily have;
2. High-frequency spatial aliasing, which arises from the necessity of sampling a continuous aperture at discrete locations through the placement of microphones.

The detrimental effects of both these problems can be minimized, and ongoing attempts to do so are the topics of a great deal of current research. Neither effect can be eliminated entirely, however, and hence must be taken into account in developing effective beamforming algorithms for real microphone arrays.

Finally, as we have shown here, the adaptive beamforming algorithms developed in the conventional literature can be effectively applied to spherical arrays. This includes the second order methods such as minimum variance distortionless response, but would also certainly include the algorithms based on the optimization of non-Gaussian and higher order statistics such as maximum kurtosis and maximum negentropy beamforming. In the view of the present authors, further investigation of these techniques in the context of spherical arrays is likely to prove one of the most promising topics of acoustic beamforming research in the coming years and decades. Moreover, the effect of this research on applications involving distant speech recognition is likely to be dramatic.

A

Modal Coefficient Calculation

Note that for integer $n, m \geq 0$ (Dunster 2010, §14.9),

$$P_n^{-m}(x) \equiv (-1)^m \frac{(n-m)!}{(n+m)!} P_n^m(x). \tag{A.1}$$

The Hankel functions of all orders can be calculated from the relation (Williams 1999, §6.4.1)

$$h_n^{(1)}(x) = (-x)^n \left(\frac{1}{x}\frac{d}{dx}\right)^n \left(\frac{e^{ix}}{ix}\right) \;\; \forall \;\; n = 0, 1, 2, \ldots. \tag{A.2}$$

Alternatively, the recurrence relations

$$h_{n+1}(x) = \frac{2n+1}{x} h_n(x) - h_{n-1}(x), \tag{A.3}$$

$$h_n'(x) = h_{n-1}(x) - \frac{n+1}{x} h_n(x). \tag{A.4}$$

can be used to determine h_n and h_n' to arbitrary order. Once the relevant Hankel function has been extracted from (A.2) or (A.3), the corresponding spherical Bessel function can be determined through the relation $h_n^{(1)} = j_n + iy_n$, where y_n is the spherical Bessel function of the second kind; both j_n and y_n are real-valued.

Williams (1999, §6.10.3) correctly give the modal coefficients defined by Meyer and Elko (2004) as

$$b_n(ka) \triangleq j_n(ka) - \frac{j_n'(ka)}{h_n'(ka)} h_n(ka), \tag{A.5}$$

where j_n is the nth spherical Bessel function and $h_n(x) = h_n^{(1)}(x)$ is the nth Hankel function of the first kind. The latter can be obtained from (Williams 1999, §6.4.1)

$$j_n(x) = (-x)^n \left(\frac{1}{x}\frac{d}{dx}\right)^n \left(\frac{\sin x}{x}\right), \tag{A.6}$$

$$h_n^{(1)} = (-x)^n \left(\frac{1}{x}\frac{d}{dx}\right)^n \left(\frac{e^{ix}}{ix}\right). \tag{A.7}$$

Mode $n = 0$: From (A.6) it follows

$$j_0(x) = \frac{\sin x}{x}, \tag{A.8}$$

such that

$$j_0'(x) = \frac{d\, j_0(x)}{dx} = \frac{x\cos x - \sin x}{x^2}. \tag{A.9}$$

Similarly, from (A.7) it follows

$$h_0(x) = h_0^{(1)}(x) = \frac{e^{ix}}{ix}, \tag{A.10}$$

so that

$$h_0'(x) = \frac{d\, h_0(x)}{dx} = \frac{(ix-1)\, e^{ix}}{x^2}. \tag{A.11}$$

Substituting (A.8–A.11) into (A.5) provides

$$b_0(ka) = \frac{\sin ka}{ka} + \frac{ka\cos ka - \sin ka}{(1-ika)ka} = \frac{\cos ka - i\sin ka}{1-ika}. \tag{A.12}$$

INDEX

References

Abhayapala TD and Ward DB 2002 Theory and design of high order sound field microphones using spherical microphone array *Proc. ICASSP*, Orlando, FL.

Arfken GB and Weber HJ 2005 *Mathematical Methods for Physicists*. Elsevier, Boston.

Askey RA and Roy R 2010 Gamma function In *NIST Handbook of Mathematical Functions* (ed. Olver FWJ, Lozier DW, Boisvert RF and Clark CW) Cambridge University Press New York, NY.

Bertsekas DP 1995 *Nonlinear Programming*. Athena Scientific, Belmont, MA, USA.

Bitzer J and Simmer KU 2001 Superdirective microphone arrays In *Microphone Arrays* (ed. Branstein M and Ward D) Springer Heidelberg pp. 19–38.

Brandstein M and Ward D 2000 Microphone arrays. *Springer*.

Carter GC 1981 Time delay estimation for passive sonar signal processing. *IEEE Trans. Acoust., Speech, Signal Processing* **ASSP-29**, 463–469.

Claesson I and Nordholm S 1992 A spatial filtering approach to robust adaptive beaming. *IEEE Trans. on Antennas and Propagation* **19**, 1093–1096.

Cohen I, Gannot S and Berdugo B 2003 An integrated real-time beamforming and postfiltering system for nonstationary noise environments. *EURASIP Journal on Applied Signal Processing* **2003**, 1064–1073.

Cox H, Zeskind RM and Owen MM 1987 Robust adaptive beamforming. *IEEE Transactions Acoustics, Speech, and Signal Processing* **ASSP-35**(10), 1365–1376.

De Haan JM, Grbic N, Claesson I and Nordholm SE 2003 Filter bank design for subband adaptive microphone arrays. *IEEE Trans. on SAP* **11**(1), 14–23.

DiBiase JH, Silverman HF and Brandstein MS 2001 Robust localization in reverberant rooms In *Microphone Arrays* (ed. Brandstein M and Ward D) Springer Verlag Heidelberg, Germany chapter 4.

Doclo S, Spriet A, Wouters J and Moonen M 2007 Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction. *Speech Communication, special issue on Speech Enhancement* **49**, 636–656.

Driscoll JR and Dennis M. Healy J 1994 Computing Fourier transforms and convolutions on the 2-sphere. *Advances in Applied Mathematics* **15**, 202–250.

Dunster TM 2010 Legendre and related functions In *NIST Handbook of Mathematical Functions* (ed. Olver FWJ, Lozier DW, Boisvert RF and Clark CW) Cambridge University Press New York, NY.

Erkelens JS, Hendriks RC, Heusdens R and Jensen J 2007 Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors. *IEEE Transactions on Audio, Speech and Language Processing* **15**, 1741–1752.

Fliege J and Maier U 1999 The distribution of points on the sphere and corresponding cubature formulae. *IMA J. Numer. Anal.* **19**, 317–334.

Gallager RG 1968 *Information Theory and Reliable Communication*. Wiley.

Gannot S and Cohen I 2004 Speech enhancement based on the general transfer function GSC and postfiltering. *IEEE Trans. on SAP* **12**, 561–571.

Gazor S and Grenier Y 1995 Criteria for positioning of sensors for a microphone array. *IEEE Trans. Speech Audio Processing* **3**, 294–303.

Gehrig T, Klee U, McDonough J, Ikbal S, Wölfel M and Fügen C 2006 Tracking and beamforming for multiple simultaneous speakers with probabilistic data association filters *Interspeech*.

Gilbert EN and Morgan SP 1955 Optimum design of antenna arrays subject to random variations. *Bell Syst. Tech. J.* **34**, 637–663.

Golub GH and Van Loan CF 1996 *Matrix Computations* third edn. The Johns Hopkins University Press, Baltimore.

Haykin S 2002 *Adaptive Filter Theory* fourth edn. Prentice Hall, New York.

Herbordt W and Kellermann W 2002 Frequency-domain integration of acoustic echo cancellation and a generalized sidelobe canceller with improved robustness. *European Transactions on Telecommunications (ETT)* **13**, 123–132.

Herbordt W and Kellermann W 2003 Adaptive beamforming for audio signal acquisition In *Adaptive Signal Processing – Applications to Real-World Problems* (ed. Benesty J and Huang Y) Springer Berlin, Germany pp. 155–194.

Herbordt W, Buchner H, Nakamura S and Kellermann W 2007 Multichannel bin-wise robust frequency-domain adaptive filtering and its application to adaptive beamforming. *IEEE Transactions on Audio, Speech and Language Processing* **15**, 1340–1351.

Hoshuyama O, Sugiyama A and Hirano A 1999 A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters. *IEEE Transactions on Signal Processing* **47**, 2677–2684.

Hyvärinen A 1999 Survey on independent component analysis. *Neural Computing Surveys* **2**, 94–128.

Hyvärinen A and Oja E 2000 Independent component analysis: Algorithms and applications. *Neural Networks* **13**, 411–430.

Kay S 1993 *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice–Hall, Englewood Cliffs, NJ.

Klee U, Gehrig T and McDonough J 2005 Kalman filters for time delay of arrival–based source localization. *Journal of Advanced Signal Processing, Special Issue on Multi–Channel Speech Processing*.

Kumatani K, Gehrig T, Mayer U, Stoimenov E, McDonough J and Wölfel M 2007 Adaptive beamforming with a minimum mutual information criterion. *IEEE Trans. on ASLP* **15**, 2527–2541.

Kumatani K, Lu L, McDonough J, Ghoshal A and Klakow D 2010a Maximum negentropy beamforming with superdirectivity *European Signal Processing Conference (EUSIPCO)*, Aalborg, Denmark.

Kumatani K, McDonough J, Klakow D, Garner PN and Li W 2008a Adaptive beamforming with a maximum negentropy criterion. *IEEE Trans. ASLP*.

Kumatani K, McDonough J, Lehman JF and Raj B 2011 Channel selection based on multichannel cross-correlation coefficients for distant speech recognition *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Edinburgh, UK.

Kumatani K, McDonough J, Rauch B and Klakow D 2010b Maximum negentropy beamforming using complex generalized gaussian distribution model *Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, Pacific Grove, CA, USA.

Kumatani K, McDonough J, Schact S, Klakow D, Garner PN and Li W 2008b Filter bank design based on minimization of individual aliasing terms for minimum mutual information subband adaptive beamforming *Proc. ICASSP*, Las Vegas, NV, USA.

Kutruff H 2009 *Room Acoustics* fifth edn. Spoon Press, New York, NY.

Li Z and Duraiswami R 2005 Hemispherical microphone arrays for sound capture and beamforming *Proc. WASPAA*, New Paltz, NY.

Li Z and Duraiswami R 2007 Flexible and optimal design of spherical microphone arrays for beamforming. *IEEE Trans. Audio Speech Lang. Proc.* **15**(2), 702–714.

Li Z, Duraiswami R, Grassi E and Davis LS 2004 Flexible layout and optimal cancellation of the orthonormality error for spherical microphone arrays *Proc. ICASSP*, Montreal, CA.

Lincoln M, McCowan I, Vepa I and Maganti HK 2005 The multi–channel Wall Street Journal audio visual corpus (mc–wsj–av): Specification and initial experiments *Proc. of ASRU*, pp. 357–362.

Marro C, Mahieux Y and Simmer KU 1998 Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. *Trans. of SAP* **6**, 240–259.

Martin R 2005 Speech enhancement based on minimum square error estimation and supergaussian priors. *IEEE Trans. Speech and Audio Processing* **13**(5), 845–856.

McCowan IA and Bourlard H 2003 Microphone array post-filter based on noise field coherence. *IEEE Trans. Speech Audio Processin* **11**, 709–716.

McDonough J, Wölfel M and Waibel A 2007 On maximum mutual information speaker–adapted training. *Computer Speech and Language*.

Meyer J and Elko GW 2002 A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield *Proc. ICASSP*, Orlando, FL.

Meyer J and Elko GW 2004 Spherical microphone arrays for 3D sound recording *Audio Signal Processing for Next–Generation Multimedia Communication Systems* Kluwer Academic Boston pp. 67–90.

Neeser FD and Massey JL 1993 Proper complex random processes with applications to information theory. *IEEE Trans. on Information Theory* **39**(4), 1293–1302.

Nordebo S, Claesson I and Nordholm S 1994 Adaptive beamforming: spatial filter designed blocking matrix. *IEEE Journal of Oceanic Engineering* **19**, 583–590.

Nordholm S, Claesson I and Bengtsson B 1993 Adaptive array noise suppression of handsfree speaker input in cars. *IEEE Trans. on Vehicular Technology* **42**, 514–518.

Olver FWJ and Maximon LC 2010 Bessel functions In *NIST Handbook of Mathematical Functions* (ed. Olver FWJ, Lozier DW, Boisvert RF and Clark CW) Cambridge University Press New York, NY.

Omologo M and Svaizer P 1994 Acoustic event localization using a crosspower–spectrum phase based technique *Proc. of ICASSP*, vol. II, pp. 273–6.

Oppenheim AV and Schafer RW 2010 *Discrete–Time Signal Processing* third edn. Prentice–Hall.

Oppenheim AV, Schafer RW and Buck JR 2009 *Discrete–Time Signal Processing* second edn. Prentice–Hall Inc.

Papoulis A and Pillai SU 2002 *Probability, Random Variables, and Stochastic Processes* fourth edn. McGraw–Hill, New York.

Roy R and Kailath T 1989 Esprit–estimation of signal parameters via rotational invariance techniques. *IEEE Transactions on Acoustics, Speech and Signal Processing* **37**, 984–995.

Sharon Gannot D, Burshtein and Weinstein E 2001 Signal enhancement using beamforming and nonstationarity withapplications to speech. *IEEE Trans. on SP* **49**, 1614–1626.

Teutsch H 2007 *Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition*. Springer, Heidelberg.

Uebel L and Woodland P 2001 Improvements in linear transform based speaker adaptation *Proc. of ICASSP*.

Vaidyanathan PP 1993 *Multirate Systems and Filter Banks*. Prentice Hall, Englewood Cliffs.

Van Trees HL 2002 *Optimum Array Processing*. Wiley, New York.

Warsitz E, Krueger A and Haeb-Umbach R 2008 Speech enhancement with a new generalized eigenvector blocking matrix for application in a generalized sidelobe canceller *Proc. of ICASSP*.
Widrow B, Duvall KM, Gooch RP and Newman WC 1982 Signal cancellation phenomena in adaptive antennas: Causes and cures. *IEEE Trans. on Antennas and Propagation* **AP-30**, 469–478.
Williams EG 1999 *Fourier Acoustics*. Academic Press, San Diego, CA, USA.
Wölfel M and McDonough J 2009 *Distant Speech Recognition*. Wiley, London.
Yan S, Sun H, Svensson UP, Ma X and Hovem JM 2011 Optimal modal beamforming for spherical microphone arrays. *IEEE Trans. on Audio Speech Language Processing* **19**(2), 361–371.