MICROPHONE ARRAY PROCESSING FOR DISTANT SPEECH RECOGNITION: FROM CLOSE-TALKING MICROPHONES TO FAR-FIELD SENSORS

Kenichi Kumatani, John McDonough, Bhiksha Raj

Distant speech recognition (DSR) holds the promise of the most natural human computer interface because it enables man-machine interactions through speech, without the necessity of donning intrusive body- or head-mounted microphones. Recognizing distant speech robustly, however, remains a challenge. This contribution provides a tutorial overview of DSR systems based on microphone arrays. In particular, we present recent work on acoustic beamforming for DSR, along with experimental results verifying the effectiveness of the various algorithms described here; beginning from a word error rate (WER) of 14.3% with a single microphone of a linear array, our state-of-the-art DSR system achieved a WER of 5.3%, which was comparable to that of 4.2% obtained with a lapel microphone. Moreover, we present an emerging technology in the area of far-field audio and speech processing based on spherical microphone arrays. Performance comparisons of spherical and linear arrays reveal that a spherical array with a diameter of 8.4 cm can provide recognition accuracy comparable or better than that obtained with a large linear array with an aperture length of 126 cm.

1. INTRODUCTION

When the signals from the individual sensors of a microphone array with a known geometry are suitably combined, the array functions as a spatial filter capable of suppressing noise, reverberation, and competing speech. Such *beamforming* techniques have received a great deal of attention within the acoustic array processing community in the recent past [1, 2, 3, 4, 5, 6, 7]. Despite this effort, however, such techniques have often been ignored within the mainstream community working on distant speech recognition. As pointed out in [6, 7], this could be due to the fact that the disparate research communities for acoustic array processing and automatic speech recognition have failed to adopt each other's best practices. For instance, the array processing community tends to ignore speaker adaptation techniques, which can compensate for mismatches between acoustic conditions during training and testing. Moreover, this community has largely preferred to work on controlled, synthetic recordings, obtained by convolving noise- and reverberation-free speech with measured, static room impulse responses, with subsequent artificial addition of noise, as in the recent PASCAL CHiME Speech Separation Challenge [8, 9, 10, 11]. A notable exception was the PASCAL Speech Separation Challenge 2 [5, 12] which featured actual array recordings of real speakers; this task, however, has fallen out of favor, to the extent that it is currently not even mentioned on the PASCAL CHiME Challenge website, nor in any of the concomitant publications. This is unfortunate because improvements obtained with novel speech enhancement techniques tend to diminish—or even disappear—after speaker adaptation; similarly, techniques that work well on artificially convolved data with artificially added noise tend to fail on data captured in real acoustic environments with real human speakers. Mainstream speech recognition researchers, on the other hand, are often unaware of advanced signal and array processing techniques. They are equally unaware of the dramatic reductions in error rate that such techniques can provide in DSR tasks.

The primary goal of this contribution is to provide a tutorial in the application of acoustic array processing to distant speech recognition that is intelligible to anyone with a general signal processing background, while still maintaining the interest of experts in the field. Our secondary goal is to bridge the gaps between the current acoustic array processing and speech recognition communities. A third and *overarching* goal is to provide a concise report on the state-of-the-art in DSR. Towards this end, we present two empirical studies: the first is a comparison of several beamforming algorithms for their effectiveness in a DSR task with real speakers in a real acoustic environment. These are conducted with a conventional linear array. The second performance comparison is between a conventional linear array and a much more compact spherical array. The latter is gaining importance as the emphasis in acoustic array processing moves from large static fixtures to smaller mobile devices such as robots.

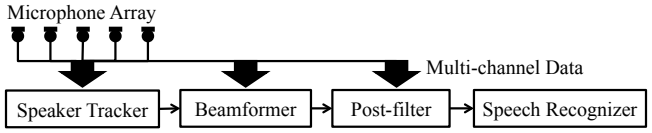


Fig. 1. Block diagram of a typical DSR system.

The remainder of this article is organized as follows. In Section 2, we provide an overview of a complete DSR system, which includes the fundamentals of array processing, speaker tracking and conventional statistical beamforming techniques. In Section 3, we consider several recently introduced techniques for beamforming with higher order statistics. This section concludes with our first set of experimental results comparing conventional beamforming techniques with those based on higher order statistics. We take up our discussion of array geometry and spherical arrays in particular in Section 4. This section also concludes with a set of experimental studies, namely that comparing the performance of a conventional linear array with a spherical array in a DSR task. In the final section of this work, we present our conclusions and plans for future work.

2. OVERVIEW OF DSR

Figure 1 shows a block diagram of a DSR system with a microphone array. The microphone array module typically consists of a speaker tracker, beamformer and post-filter. The speaker tracker estimates a speaker's position. Given that position estimate, the beamformer emphasizes sound waves coming from the direction of interest or "look direction". The beamformed signal can be further enhanced with post-filtering. The final output is then fed into a speech recognizer. We note that this framework can readily incorporate other information sources such as a mouth locator based on video data [13].

2.1. Fundamental Issues in Microphone Array Processing

As shown in Figure 2, the array processing components of a DSR system are prone to several errors. Firstly, there are errors in speaker tracking which cause the beam to be "steered" in the wrong direction [14]; such errors can in turn cause signal cancellation. Secondly, the individual microphones in the array can have different amplitude and phase responses even if they are of the same type [15, §5.5]. Fi-

nally, the placement of the sensors can deviate from their nominal positions. All of these factors degrade beamforming performance.

2.2. Speaker Tracking

The speaker tracking problem is generally distinguished from the speaker localization problem. Speaker localization methods estimate a speaker's position at a single instant in time without relying on past information. On the other hand, speaker tracking algorithms consider a trajectory of instantaneous position estimates.

Speaker localization techniques could be categorized into three approaches: seeking a position which provides the maximum steered response power (SRP) of a beamformer [16, §8.2.1], localizing a source based on the application of high-resolution spectral estimation techniques such as subspace algorithms [17, §9.3], and estimating sources' positions from time delays of arrival (TDOA) at the microphones. Due to computational efficiency as well as robustness against mismatches of signal models and microphone errors, TDOA-based speaker localization approaches are perhaps the most popular in DSR. Here, we briefly introduce speaker tracking methods based on the TDOA.

Shown in Figure 3a is a sound wave propagating from a point $\mathbf{x}$ to each microphone located at $\mathbf{m}_s$ for all $s = 0, \cdots, S-1$ where S is the total number of sensors. Assuming that the position of each microphone is specified in Cartesian coordinates, denote the distance between the point source and each microphone as $D_s \triangleq \|\mathbf{x} - \mathbf{m}_s\| \ \forall \ s = 0, \cdots, S-1$. Then, the TDOA between microphones m and n can be expressed as

$$\tau_{m,n}(\mathbf{x}) \triangleq (D_m - D_n)/c, \tag{1}$$

where c is the speed of sound. Notice that equation (1) implies that the *wavefront*—a surface comprised of the locus of all points on the same phase—is spherical.

In the case that the array is located far from the speaker, the wavefront can be assumed to be planar, which is called the far-field assumption. Figure 3b illustrates a plane wave propagating from the far-field to the microphones. Under the far-field assumption, the TDOA becomes a function of the angle θ between the *direction of arrival* (DOA) and the line connecting two sensors' positions, and equation (1) can be simplified as

$$\tau_{m,n}(\theta) \triangleq d_{m,n} \cos\theta / c, \tag{2}$$

where $d_{m,n}$ is the distance between the microphones m and n.

Various techniques have been developed for estimation of the TDOAs. A comprehensive overview of those algorithms is provided by [18] and comparative studies on real data can be found in [19]. From the TDOA between the microphone pairs, the speaker's position can be computed using classical methods, namely, spherical intersection, spherical interpolation or linear intersection [2, §10.1]. These methods can readily be extended to track a moving speaker by applying a Kalman filter (KF) to smooth the time series of the instantaneous estimates as in [16, §10]. Klee et al. [20] demonstrated,

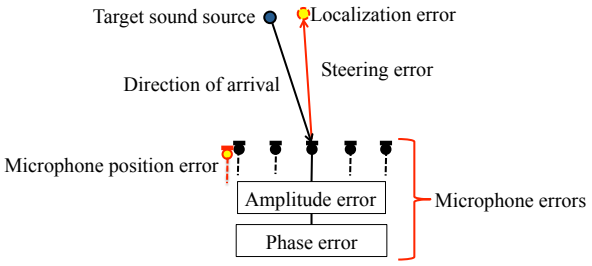


Fig. 2. Representative errors in microphone array processing.

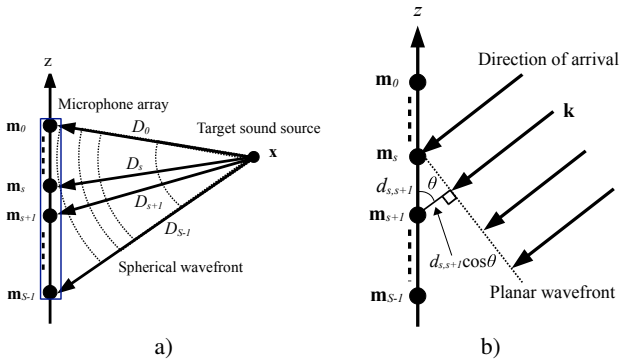


Fig. 3. Propagation of a) the spherical wave and b) plane wave.

however, that instead of smoothing a series of instantaneous position estimates, better tracking could be performed by simply using the TDOAs as a sequence of observations for an extended Kalman filter (EKF) and estimating the speaker's position directly from the standard EKF state estimate update formulae. Klee's algorithm was extended to incorporate video features in [21], and to track multiple simultaneous speakers [22].

2.3. Conventional Beamforming Techniques

In the case of the spherical wavefront depicted in Figure 3a, let us define the *propagation delay* as $\tau_s \triangleq D_s/c$. In the far-field case shown in Figure 3b, let us define the *wavenumber* $\mathbf{k}$ as a vector perpendicular to the planar wavefront pointing in the direction of propagation with magnitude $\omega/c = 2\pi/\lambda$. Then, the propagation delay with respect to the origin of the coordinate system for microphone s is determined through $\omega\tau_s = \mathbf{k}^T\mathbf{m}_s$. The simplest model of wave propagation assumes that a signal $f(t)$, carried on a plane wave, reaches all sensors in an array, but not at the same time. Hence, let us form the vector

$$\mathbf{f}(t) = [f(t-\tau_0) \quad f(t-\tau_1) \quad \cdots \quad f(t-\tau_{S-1})]^T$$

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

$$\mathbf{v}(\mathbf{k},\omega) \triangleq \left[e^{-i\omega\tau_0} \quad e^{-i\omega\tau_1} \quad \cdots \quad e^{-i\omega\tau_{S-1}}\right]^T \tag{3}$$

is the *array manifold vector*. The latter is manifestly a vector of phase delays for a plane wave with wavenumber $\mathbf{k}$. To a first order, the array manifold vector is a complete description of the interaction of a propagating wave and an array of sensors.

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

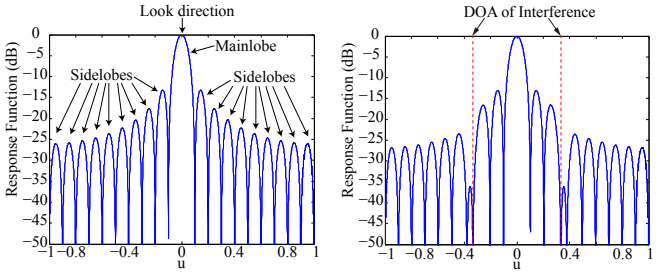
$$Y(\omega) = \mathbf{w}^H(\omega)\,\mathbf{X}(\omega), \tag{4}$$

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

$$\mathbf{w}_{\text{DS}} \triangleq \mathbf{v}(\mathbf{k},\omega)/S. \tag{5}$$

Substituting $\mathbf{X}(\omega) = \mathbf{F}(\omega) = F(\omega)\mathbf{v}(\mathbf{k},\omega)$ into (4) provides

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



a) b)

Fig. 4. Beampatterns of a) the delay-and-sum beamformer and b) MVDR beamformer as a function of $u = \cos\theta$ for the linear array; the noise covariance matrix of the MVDR beamformer is computed with the interference plane waves propagating from $u = \pm 1/3$.

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

$$\mathbf{w}^H(\omega)\,\mathbf{v}(\mathbf{k},\omega) = 1. \tag{6}$$

Hereafter we will say that any weight vector $\mathbf{w}(\omega)$ achieving (6) satisfies the *distortionless constraint*, which implies that any wave impinging from the look direction is neither amplified nor attenuated.

Figure 4a shows the *beampattern* of the DS beamformer, which indicates the sensitivity of the beamformer in decibels to plane waves impinging from various directions. The beampatterns are plotted as a function of $u = \cos\theta$ where θ is the angle between the DOA and the axis of the linear array. The beampatterns in Figure 4 were computed for a linear array of 20 uniformly-spaced microphones with an intersensor spacing of $d = \lambda/2$, where λ is the wavelength of the impinging plane waves; the look direction is $u = 0$. The lobe around the look direction is the *mainlobe*, while the other lobes are *sidelobes*. The large sidelobes indicate that the suppression of noise and interference off the look direction is poor; in the case of DS beamforming, the first sidelobe is only 13 dB below the mainlobe.

To improve upon noise suppression performance provided by the DS beamformer, it is possible to adaptively suppress spatially-correlated noise and interference $\mathbf{N}(\omega)$, which can be achieved by adjusting the weights of a beamformer so as to minimize the variance of the noise and interference at the output subject to the distortionless constraint (6). More concretely, we seek $\mathbf{w}(\omega)$ achieving

$$\text{argmin}_{\mathbf{W}}\ \mathbf{w}^H(\omega)\,\mathbf{\Sigma_N}(\omega)\,\mathbf{w}(\omega), \tag{7}$$

subject to (6), where $\mathbf{\Sigma_N} \triangleq \mathcal{E}\{\mathbf{N}(\omega)\mathbf{N}^H(\omega)\}$ and $\mathcal{E}\{\cdot\}$ is the expectation operator. In practice, $\mathbf{\Sigma_N}$ is computed by averaging or recursively updates the noise covariance matrix [17, §7]. The weight vectors obtained under these conditions correspond to the *minimum variance distortionless response* (MVDR) beamformer, which has the well-known solution [2, §13.3.1]

$$\mathbf{w}^H_{\text{MVDR}}(\omega) = \frac{\mathbf{v}^H(\mathbf{k},\omega)\,\mathbf{\Sigma}_\mathbf{N}^{-1}(\omega)}{\mathbf{v}^H(\mathbf{k},\omega)\,\mathbf{\Sigma}_\mathbf{N}^{-1}(\omega)\,\mathbf{v}(\mathbf{k},\omega)}. \tag{8}$$

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

Figure 4b shows the beampattern of the MVDR beamformer for the case of two plane wave interferers arriving from directions $u = \pm 1/3$. It is apparent from the figure that such a beamformer can place deep nulls on the interference signals while maintaining unity gain in the look direction. In the case of $\mathbf{\Sigma_N} = \mathbf{I}$, which indicates that the noise field is spatially-uncorrelated, the MVDR and DS beamformers are equivalent.

Depending on the acoustic environment, adapting the sensor weights $\mathbf{w}(\omega)$ to suppress discrete sources of interference can lead to excessively large sidelobes, resulting in poor system robustness. A simple technique for avoiding this is to impose a quadratic constraint $\|\mathbf{w}\|^2 \leq \gamma$, for some $\gamma > 0$, in addition to the distortionless constraint (6), when estimating the sensor weights. The MVDR solution will then take the form [2, §13.3.7]

$$\mathbf{w}^H_{\text{DL}} = \frac{\mathbf{v}^H\left(\mathbf{\Sigma_N} + \sigma_\text{d}^2\mathbf{I}\right)^{-1}}{\mathbf{v}^H\left(\mathbf{\Sigma_N} + \sigma_\text{d}^2\mathbf{I}\right)^{-1}\mathbf{v}}, \tag{9}$$

which is referred to as *diagonal loading* where σ_d^2 is the loading level; the dependence on ω in (9) has been suppressed for convenience. While (9) is straightforward to implement, there is no direct relationship between γ and σ_d^2; hence the latter is typically set either based on experimentation or through an iterative procedure. Increasing σ_d^2 decreases $\|\mathbf{w}_{\text{DL}}\|$, which implies that the *white noise gain* (WNG) also increases [23]; WNG is a measure of the robustness of the system to the types of errors shown in Figure 2.

A theoretical model of diffuse noise that works well in practice is the spherically isotropic field, wherein spatially separated microphones receive equal energy and random phase noise signals from all directions simultaneously [16, §4]. The MVDR beamformer with the diffuse noise model is called the *super-directive beamformer* [2, §13.3.4]. The super-directive beamforming design is obtained by replacing the noise covariance matrix $\mathbf{\Sigma_N}(\omega)$ with the coherence matrix $\mathbf{\Gamma}(\omega)$ whose (m,n)-th component is given by

$$\Gamma_{m,n}(\omega) = \text{sinc}\left(\frac{\omega d_{m,n}}{c}\right), \tag{10}$$

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

$$\mathbf{\Sigma_N}(\omega) = \Sigma_N(\omega)\mathbf{v}(\mathbf{k}_\text{I})\mathbf{v}^H(\mathbf{k}_\text{I}) + \sigma_\text{SI}^2\mathbf{\Gamma}(\omega), \tag{11}$$

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

The MVDR beamformer is of particular interest because it forms the preprocessing component of two other important beamforming structures. Firstly, the MVDR beamformer followed by a suitable post-filter yields the *maximum signal-to-noise ratio* beamformer [17, §6.2.3]. Secondly, and more importantly, by placing a Wiener filter [24, §2.2] on the output of the MVDR beamformer, the *minimum mean-square error* (MMSE) beamformer is obtained [17, §6.2.2]. Such *post-filters* are important because it has been shown that they can yield significant reductions in error rate [25, 5]. Of the several post-filtering methods proposed in the literature [26], the *Zelinski post-filtering* [27] technique is arguably the simplest practical implementation of a Wiener filter. Wiener filters in their pure form are unrealizable because they assume that the spectrum of the desired signal is available. The Zelinski post-filtering method uses the auto- and cross-power spectra of the multi-channel input signals to estimate the target signal and noise power spectra effectively under the assumption of zero cross-correlation between the noises at different sensors. We have employed the Zelinski post-filter for the experiments described in Sections 3.4 and 4.3.

The MVDR beamformer can be implemented in generalized sidelobe canceller (GSC) configuration [17, §6.7.3] as shown in Figure 5. For the input snapshot vector $\mathbf{X}(t)$ at a frame t, the output of a GSC beamformer can be expressed as

$$Y(t) = [\mathbf{w}_\text{q}(t) - \mathbf{B}(t)\mathbf{w}_\text{a}(t)]^H\,\mathbf{X}(t), \tag{12}$$

where $\mathbf{w}_\text{q}$ is the *quiescent weight vector*, $\mathbf{B}$ is the *blocking matrix*, and $\mathbf{w}_\text{a}$ is the *active weight vector*. In keeping with the GSC formalism, $\mathbf{w}_\text{q}$ is chosen to satisfy the distortionless constraint (6) [2,

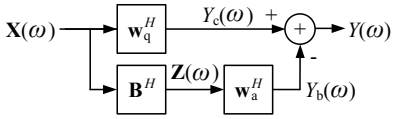


Fig. 5. Generalized sidelobe cancellation beamformer in the frequency domain.

§13.6]. The blocking matrix $\mathbf{B}$ is chosen to be orthogonal to $\mathbf{w}_{\mathrm{q}}$, such that $\mathbf{B}^H \mathbf{w}_{\mathrm{q}} = \mathbf{0}$. This orthogonality implies that the distortionless constraint will be satisfied for any choice of $\mathbf{w}_{\mathrm{a}}$.

The MVDR beamformer and its variants can effectively suppress sources of interference. They can also potentially cancel the target signal, however, in cases wherein signals correlated with the target signal arrive from directions other than the look direction. This is precisely what happens in all real acoustic environments due to reflections from hard surfaces such as tables, walls and floors. A brief overview of techniques for preventing signal cancellation can be found in [28].

For the empirical studies reported in Section 3.4 and 4.3, subband analysis and synthesis were performed with a uniform DFT filter bank based on the modulation of a single prototype impulse response [2, §11][29]. Subband adaptive filtering can reduce the computational complexity associated with time domain adaptive filters and improve convergence rate in estimating filter coefficients. The complete processing chain—including subband analysis, beamforming, and subband synthesis—is shown in Figure 6; briefly it comprises the steps of subband filtering as indicated by the blocks labeled $H(\omega_m)$ followed by decimation. Thereafter the decimated subband samples $\mathbf{X}(\omega_m)$ are weighted and summed during the beamforming stage. Finally, the beamformed samples are expanded and processed by a synthesis filter $G(\omega_m)$ to obtain a time-domain signal. In order to alleviate the unwanted aliasing effect caused by arbitrary magnitude scaling and phase shifts in adaptive processing, our analysis and synthesis filter prototype [29] is designed to minimize individual aliasing terms separately instead of maintaining the perfect reconstruction property.

Working in the discrete time and discrete frequency domains requires that the definition (3) of the array manifold vector be modified as

$$\mathbf{v}(\mathbf{k}, \omega_m) \triangleq \left[e^{-i\omega_m \tau_0 f_{\mathrm{s}}} \quad e^{-i\omega_m \tau_1 f_{\mathrm{s}}} \ldots \quad e^{-i\omega_m \tau_{S-1} f_{\mathrm{s}}} \right],$$

where f_{s} is the digital sampling frequency, and $\omega_m = 2\pi m/M$ for all $m = 0, \ldots, M-1$ are the subband center frequencies.

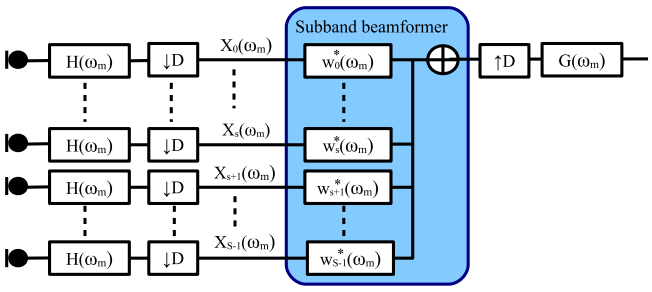


Fig. 6. Schematic view for beamforming in the subband domain.

3. BEAMFORMING WITH HIGHER-ORDER STATISTICS

The conventional beamforming algorithms estimate their weights based on the covariance matrix of the snapshot vectors. In other words, the conventional beamformer's weights are determined solely by second-order statistics (SOS). Beamforming with higher-order statistics (HOS) has recently been proposed in the literature [28]; it has been demonstrated that such HOS beamformers do *not* suffer from signal cancellation.

In this section, we introduce a concept of non-Gaussianity and describe how the fine the structure of a non-Gaussian *probability density function* (pdf) can be characterized by measures such as kurtosis and negentropy. Moreover, we present empirical evidence that speech signals are highly non-Gaussian; thereafter we discuss beamforming algorithms based on maximizing non-Gaussian optimization criteria.

3.1. Motivation for Maximizing non-Gaussianity

The *central limit theorem* [30] 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. Hence, a desired signal corrupted with statistically independent noise will clearly be closer to Gaussian than the original clean signal. When a non-stationary signal such as speech is corrupted with the reverberation, portions of the speech that are essentially independent—given that the room reverberation time (300-500 ms) is typically much longer than the duration of any phone (100 ms)—segments of an utterance that are essentially independent will be added together. This implies that the reverberant speech must similarly be closer to Gaussian than the original "dry" signal. Hence, by attempting to restore the original super-Gaussian statistical characteristics of speech, we can expect to ameliorate the deleterious effects of *both* noise and reverberation.

There are several popular criteria for measuring a degree of non-Gaussianity. Here, we review *kurtosis* and *negentropy* [30, §8].

Kurtosis Among several definitions of kurtosis for an RV Y with zero mean, the kurtosis measure we adopt here is

$$\mathrm{kurt}(Y) \triangleq \mathcal{E}\{|Y|^4\} - \beta_{\mathrm{K}} (\mathcal{E}\{|Y|^2\})^2. \qquad (13)$$

where β_{K} is a positive constant, which is typically set to $\beta_{\mathrm{K}} = 3$ for kurtosis of real-valued RVs in order to ensure that the Gaussian has zero kurtosis; pdfs with positive kurtosis are super-Gaussian, and those with negative kurtosis are sub-Gaussian. An empirical estimate of kurtosis can be computed given some samples from the output of a beamformer by replacing the expectation operator of (13) with a time average.

Negentropy Entropy is the basic measure for information in *information theory* [30]. The differential entropy for continuous complex-valued RVs Y with the pdf $p_Y(\cdot)$ is defined as

$$H(Y) \triangleq -\int p_Y(v) \log p_Y(v) dv = -\mathcal{E}\left\{\log p_Y(v)\right\}. \qquad (14)$$

Another criterion for measuring the degree of super-Gaussianity is negentropy, which is defined as

$$\mathrm{neg}(Y) \triangleq H(Y_{\mathrm{gauss}}) - H(Y), \qquad (15)$$

where Y_{gauss} is a Gaussian variable with the same variance σ_Y^2 as Y. For complex-valued RVs, the entropy of Y_{gauss} can be expressed as

$$H(Y_{\mathrm{gauss}}) = \log \left|\sigma_Y^2\right| + (1 + \log \pi). \qquad (16)$$

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 from the Gaussian pdf. Computing the entropy $H(Y)$

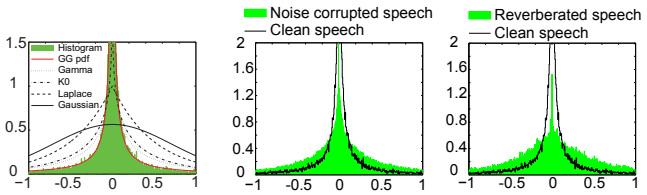


Fig. 7. Histograms of real parts of subband frequency components of clean speech and a) pdfs, b) noise-corrupted speech and c) reverberated speech.

of a super-Gaussian variable Y requires knowledge of its specific pdf. Thus, it is important to find a family of pdfs capable of closely modeling the distributions of actual speech signals. The generalized Gaussian pdf (GG-pdf) is frequently encountered in the field of *independent component analysis* (ICA). Accordingly, we used the GG-pdf for the DSR experiments described in Section 3.4. The form of the GG-pdf and entropy is described in [31]. As with kurtosis, an empirical version of entropy can be calculated by replacing the ensemble expectation with a time average over samples of the beamformer's output.

As indicated in (13), the kurtosis measure considers not only the variance but also the fourth moment, a higher-order statistic. Hence, empirical estimates of kurtosis can be strongly influenced by a few samples with a low observation probability, or *outliers*. Empirical estimates of negentropy are generally more robust in the presence of outliers than those for kurtosis [30, §8].

Distribution of Speech Samples Figure 7a shows a histogram of the real parts of subband samples at frequency 800 Hz computed from clean speech. Figure 7a also shows the Gaussian distribution and several super-Gaussian pdfs: Laplace, K_0, Gamma and GG-pdf trained with the actual samples. As shown in the figure, the super-Gaussian pdfs are characterized by a spikey peak at the mean and heavy tails in regions well-removed from the mean; it is clear that the pdf of the subband samples of clean speech is super-Gaussian. It is also clear from Figures 7b and 7c that the distributions of the subband samples corrupted with noise and reverberation get closer to the Gaussian pdf. These results suggest that the effects of noise and reverberation can be suppressed by adjusting beamformer's weights so as to make the distribution of its outputs closer to that of clean speech, that is, the super-Gaussian pdf.

3.2. Beamforming with the maximum super-Gaussian criterion

Given the GSC beamformer's output Y, we can obtain a measure of its super-Gaussianity with (13) or (15). Then, we can adjust the active weight vector so as to achieve the maximum kurtosis or negentropy while maintaining the distortionless constraint (6) under the formalism of the GSC. In order to avoid a large active weight vector, a regularization term is added, which has the same function as diagonal loading in conventional beamforming. We denote the cost function as

$$\mathcal{J}(Y) = J(Y) - \alpha \|\mathbf{w}_{\mathrm{a}}\|^2 \qquad (17)$$

where $J(Y)$ is the kurtosis or negentropy, and $\alpha > 0$ is a constant. A stricter constraint on the active weight vector can also be imposed as $\|\mathbf{w}_{\mathrm{a}}\|^2 \leq \gamma$ for some real $\gamma > 0$. Due to the absence of a closed-form solution for that $\mathbf{w}_{\mathrm{a}}$ maximizing (17), we must resort to a numerical optimization algorithm; details can be found in [28, 31] for maximum negentropy (MN) beamforming and [32] for maximum kurtosis (MK) beamforming.

As shown in [28], beamforming algorithms based on the maximum super-Gaussian criteria attempt to strengthen the reflected wave of a desired source so as to enhance speech. Of course, any reflected signal would be delayed with respect to the direct path signal. Such a delay would, however, manifest itself as a phase shift in the frequency domain if the delay is shorter than the length of the analysis filter, and could thus be removed through a suitable choice of $\mathbf{w}_{\mathrm{a}}$ based on the maximum super-Gaussian criteria. Hence, the MN and MK beamformers offer the possibility of suppressing the reverberation effect by compensating the delays of the reflected signals.

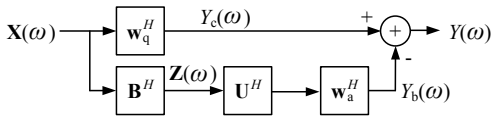


Fig. 8. Maximum kurtosis beamformer with the subspace filter.

In real acoustic environments, the desired signal will arrive from many directions in addition to the direct path. Therefore, it is not feasible for conventional adaptive beamformers to avoid the signal cancellation effect, as demonstrated in experiments described later. On the other hand, MN or MK beamforming can estimate the active weight vector to enhance target speech without signal cancellation solely based on the super-Gaussian criterion.

3.3. Online Implementation with Subspace Filtering

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 can become acute because the optimization surfaces encountered in HOS beamforming are less regular than those in conventional beamforming. In order to achieve efficient estimation, an eigen- or subspace filter [17, §6.8] can be used as a pre-processing step for estimation of the active weight vector. In this section, we review MK beamforming with subspace filtering, which was originally proposed in [33].

Figure 8 shows configuration of the MK beamformer with the subspace filter. The beamformer's output can be expressed as

$$Y(t) = [\mathbf{w}_{\mathrm{q}}(t) - \mathbf{B}(t)\mathbf{U}(t)\mathbf{w}_{\mathrm{a}}(t)]^H \mathbf{X}(t). \qquad (18)$$

The difference between (12) and (18) is the subspace filter between the blocking matrix and active weight vector. The 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 a decomposition can be achieved by performing an eigendecomposition on the covariance matrix of the output of the blocking matrix. Then, we select the eigenvectors corresponding to the largest eigenvalues as the dominant modes [17, §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. In doing so, we can readily subtract the averaged ambient noise component from the beamformer's output. Moreover, the dimensionality reduction of the active weight vector leads to computationally efficient and reliable estimation.

Figure 9 illustrates actual eigenvalues sorted in descending order over frequencies. In order to generate the plots of the figure, we computed the eigenvalues from the outputs of the blocking matrix on the real data described in [34]. As shown in Figure 9, there is a distinct difference between the small and large eigenvalues at each frequency bin. Thus, it is relatively easy to determine the number of the dominant eigenvalues D especially in the case where the number of the microphones is much larger than the number of the spatially-correlated signals.

Based on equation (13) and (18), the kurtosis of the outputs is computed from an incoming block of input subband samples instead of using the entire utterance. We incrementally update the dominant modes and active weight vector at each block of samples. Again,

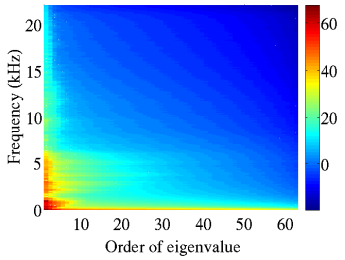


Fig. 9. Eigenvalues sorted in descending order over frequencies.

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
Super-directive (SD) BF	71.4	31.9	16.6	14.1
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

Table 1. Word error rates for each beamforming algorithm after every decoding pass.

must to resort to a gradient-based optimization algorithm for estimation of the active weight vector. The gradient is iteratively calculated with a block of subband samples until the kurtosis value of the beamformer's outputs converges. 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.

3.4. Evaluation of Beamforming Algorithms

In this section, we compare the SOS-based beamforming methods to the HOS-based algorithms. The results of DSR experiments reported here were obtained on speech material from the Multi-Channel Wall Street Journal Audio Visual Corpus (MC-WSJ-AV); see [4] 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 diameters of 20 cm, although we used only one of these arrays for our experiments. Additionally, each speaker was equipped with a *close talking microphone* (CTM) to provide the best possible 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 fixed positions. The test data set 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 39.2 minutes of speech.

Prior to beamforming, we first estimated the speaker's position with the tracking system described in [22]. Based on an average speaker position estimated for each utterance, active weight vectors $\mathbf{w}_a$ were estimated for the source on a per utterance basis.

Four decoding passes were performed on waveforms obtained with various beamforming algorithms. The details of the feature extraction component of our ASR system are given in [28]. 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. A description of the four decoding passes follows: (1) decode with the unadapted, conventional *maximum likelihood* (ML) acoustic model; (2) estimate *vocal tract length normalization* (VTLN) [2, §9] and *constrained maximum likelihood linear regression* (CMLLR) parameters [2, §9] for each speaker, then redecode with the conventional ML acoustic model; (3) estimate VTLN, CMLLR and *maximum likelihood linear regression* (MLLR) [2, §9] parameters, then redecode with the conventional model; and (4) estimate VTLN, CMLLR and MLLR parameters for each speaker, then redecode with the ML-SAT model [2, §8.1]. The standard WSJ trigram language model was used in all passes of decoding.

Table 1 shows the word error rates (WERs) for each beamforming algorithm. As references, WERs are also reported for the CTM and single array channel (SAC). It is clear from Table 1 that dramatic improvements in recognition performance are achieved by the speaker adaptation techniques which are also able to adapt the acoustic models to the noisy acoustic environment. Although the use of speaker adaptation techniques can greatly reduce WER, they often also reduce the improvement provided by signal enhancement techniques. As speaker adaptation is integral to the state-of-the-art, it is essential to report WER all such techniques having been applied; unfortunately this is rarely done in the acoustic array processing literature. It is also 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 the super-directive beamformer (SD BF) [2, §13.3.4], the MVDR beamformer (MVDR BF) and the generalized eigenvector beamformer (GEV BF) [35]. This is because the HOS-based beamformers can use the echos of the desired signal to enhance the final output of the beamformer and, as mentioned in Section 3.2, do not suffer from signal cancellation. Unlike the SOS beamformers, the HOS beamformers perform best when the active weight vector is adapted while the desired speaker is active. The SOS-based and HOS-based beamformers can be profitably combined because they employ different criteria for estimation of the active weight vector. For example, the super-directive beamformer's weight can be used as the quiescent weight vector in GSC configuration [36]. We observe from Table 1 that the maximum negentropy beamformer with super-directive beamformer (SD MN BF) provided the best recognition performance in this task.

3.5. Effect of HOS-based Beamforming with Subspace Filtering

In this section, we investigate effects of MK beamforming with subspace filtering. The *Copycat* data [34] were used as test material here. The speech material in this corpus was captured with the 64-channel linear microphone array; the sensors were arranged linearly with a 2 cm inter-sensor spacing. In order to provide a reference, subjects were also equipped with lapel microphones with a wireless connection to a preamp input. All the audio data were stored at 44.1 kHz with a 16 bit resolution. The test set consists of 356 (1,305 words) utterances spoken by an adult and 354 phrases (1,297 words) uttered by nine children who aged four to six. The vocabulary size is 147 words. As is typical for children in this age group, pronunciation was quite variable and the words themselves were sometimes indistinct.

For this task, the acoustic models were trained with two publicly available corpora of children's speech, the Carnegie Mellon University (CMU) Kids' corpus and the Center for Speech and Language Understanding (CSLU) Kids' corpus. The details of the ASR system are described in [33]. The decoder used here consists of three passes; the first and second passes are the same as the ones described in Section 3.4 but the third pass includes processing of the third and fourth passes described in Section 3.4.

Table 2 shows word error rates (WERs) of every decoding pass obtained with the single array channel (SAC), super-directive beamforming (SD BF), conventional maximum kurtosis beamforming

Algorithm	Pass (%WER)					
	1		2		3	
	Exp.	Child	Exp.	Child	Exp.	Child
SAC	9.2	31.0	3.8	17.8	3.4	14.2
SD BF	5.4	24.4	2.5	9.6	2.2	7.6
MK BF	5.4	25.1	2.5	9.0	2.1	6.5
MK BF w SF	6.3	25.4	1.2	7.4	0.6	5.3
CTM	3.0	12.5	2.0	5.7	1.9	4.2

Table 2. Word error rates (WERs) for each decoding pass.

Algorithm	Block size (second)	Pass (%WER)			
		2		3	
		Exp.	Child	Exp.	Child
Conventional MK BF	0.25	4.4	15.8	3.5	12.0
	0.5	3.4	9.2	3.1	7.3
	1.0	2.4	10.3	2.2	6.9
	2.5	2.5	9.0	2.1	6.5
MK BF w SF	0.25	2.5	14.1	1.5	9.7
	0.5	1.3	8.7	1.0	7.0
	1.0	1.2	7.4	0.6	5.3

Table 3. WERs as a function of amounts of adaptation data.

(MK BF) and maximum kurtosis beamforming with the subspace filter (MK BF w SF). The WERs obtained with the lapel microphone are also provided as a reference. It is also clear from Table 2 that the maximum kurtosis beamformer with subspace filtering achieved the best recognition performance.

Table 3 shows the WERs of the conventional and new MK beamforming algorithms as a function of amounts of adaptation data in each block. We can see from Table 3 that MK beamforming with subspace filtering (MK BF w SF) provides better recognition performance with the same amount of the data than conventional MK beamforming. In the case that little adaptation data is available, the MK beamforming does not always improve the recognition performance due to the dependency of the initial value and noisy gradient information which can significantly change over the blocks. The results in Table 3 suggest that unreliable estimation of the active weight vector can be avoided by constraining the search space with a subspace filter, as described in Section 3.3. Note that the solution of the eigendecomposition does not depend on the initial value in contrast to the gradient-based numerical optimization algorithm.

4. EFFECTS OF ARRAY GEOMETRY

In the majority of the DSR literature, results obtained with linear or circular arrays have been reported. On the other hand, in the field of acoustic array processing, spherical microphone array techniques have recently received a great deal of attention [37, 38, 39, 40]. The advantage of spherical arrays is that they can be pointed at a desired speaker in any direction with equal effect; the shape of the beampattern is invariant to the look direction. The following sections provide a review of beamforming methods in the spherical harmonics domain. Thereafter we provide a comparison of spherical and linear arrays in terms of DSR performance.

4.1. Spherical Microphone Arrays

In this section, we describe how beamforming is performed in the spherical harmonics domain. We will use the spherical coordinate system (r, θ, ϕ) shown in Figure 10 and denote the pair of *polar angle* θ and *azimuth* ϕ as $\Omega = (\theta, \phi)$.

Spherical Harmonics Let us begin by defining the *spherical harmonic* of order n and degree m [37] as

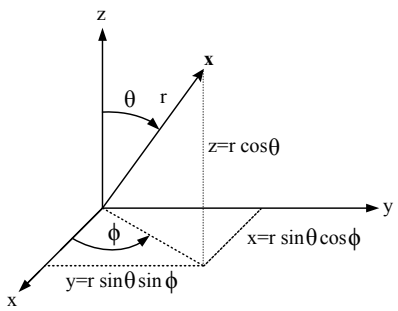


Fig. 10. Relationship between the Cartesian and spherical coordinate systems.

$$Y_n^m(\Omega) \triangleq \sqrt{\frac{(2n+1)}{4\pi}\frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta) e^{im\phi}, \qquad (19)$$

where $P_n^m(\cdot)$ denotes the associated Legendre function [41, §6.10.1]. Figure 11 shows the magnitude for the spherical harmonics, $Y_0 \triangleq Y_0^0$, $Y_1 \triangleq Y_1^0$, $Y_2 \triangleq Y_2^0$ and $Y_3 \triangleq Y_3^0$ in three-dimensional space. The spherical harmonics satisfy the *orthonormality condition* [37, 38],

$$\delta_{n,n'}\,\delta_{m,m'} = \int_\Omega Y_{n'}^{m'}(\Omega)\bar{Y}_n^m(\Omega)d\Omega \qquad (20)$$

$$= \int_0^{2\pi}\int_0^{\pi} Y_{n'}^{m'}(\theta,\phi)\bar{Y}_n^m(\theta,\phi)\sin\theta d\theta\, d\phi, \qquad (21)$$

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

Spherical Fourier Transform In Section 2.3, we defined the wavenumber as a vector perpendicular to the front of a plane wave of frequency ω pointing in the direction of propagation with a magnitude of ω/c. Now let us define the *wavenumber scalar* as $k = |\mathbf{k}| = \omega/c$; when no confusion can arise, we will also refer to k as simply the wavenumber. Let us assume that a plane wave of wavenumber k with unit power is impinging on a rigid sphere of radius a from direction $\Omega_0 = (\theta_0, \phi_0)$. The total complex sound pressure on the sphere surface at Ω_s can be expressed as

$$G(ka, \Omega_s, \Omega_0) = 4\pi \sum_{n=0}^{\infty} i^n b_n(ka) \sum_{m=-n}^{n} \bar{Y}_n^m(\Omega_0) Y_n^m(\Omega_s), \quad (22)$$

where the *modal coefficient* $b_n(ka)$ is defined as [37, 39]

$$b_n(ka) \triangleq j_n(ka) - \frac{j_n'(ka)}{h_n'(ka)} h_n(ka); \qquad (23)$$

j_n and h_n are the spherical Bessel function of the first kind and the Hankel function of the first kind [42, §10.2], respectively, and a prime

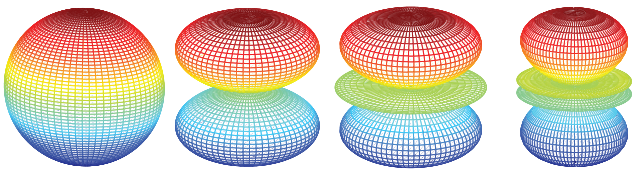


Fig. 11. Magnitude of spherical harmonics, a) Y_0, b) Y_1, c) Y_2 and d) Y_3.

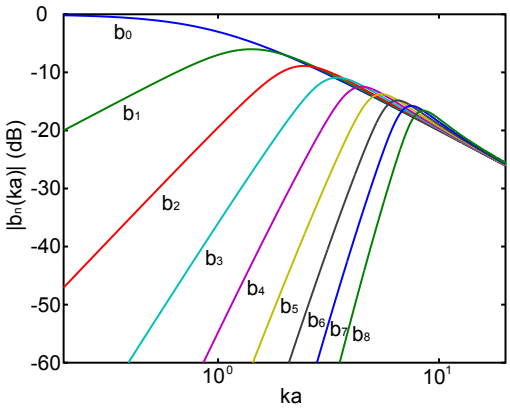


Fig. 12. Magnitude of the modal coefficients as a function of ka.

indicates the derivative of a function with respect to its argument. Figure 12 shows the magnitude of the modal coefficients as a function of ka. It is apparent from the figure that the spherical array will have poor directivity at the lowest frequencies—such as $ka = 0.2$ which corresponds to 260 Hz for $a = 4.2$ cm—inasmuch as only Y_0 is available for beamforming; amplifying the higher order modes at these frequencies would introduce a great deal of sensor self noise into the beamformer output. From Figure 11 a), however, it is clear that Y_0 is completely isotropic; i.e., it has no directional characteristics and hence provides no improvement in directivity over a single omnidirectional microphone.

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

$$G_n^m(ka, \Omega_0) = \int_\Omega G(ka, \Omega, \Omega_0) \bar{Y}_n^m(\Omega) d\Omega \tag{24}$$

and the inverse transform is defined as

$$G(ka, \Omega, \Omega_0) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} G_n^m(ka, \Omega_0) Y_n^m(\Omega). \tag{25}$$

The transform (24) can be intuitively interpreted as the decomposition of the sound field into the spherical harmonics illustrated in Figure 11.

Upon substituting the plane wave (22) into (24), we can represent the plane wave in the spherical harmonics domain as

$$G_n^m(ka, \Omega_0) = 4\pi\, i^n\, b_n(ka)\, \bar{Y}_n^m(\Omega_0). \tag{26}$$

In order to understand how beamforming may be performed in the spherical harmonic domain, we need only define the *modal array manifold vector* [43, §5.1.2] as

$$\mathbf{v}(ka, \Omega_0) \triangleq \begin{bmatrix} G_0^0(ka, \Omega_0) \\ G_1^{-1}(ka, \Omega_0) \\ G_1^0(ka, \Omega_0) \\ G_1^1(ka, \Omega_0) \\ G_2^{-2}(ka, \Omega_0) \\ G_2^{-1}(ka, \Omega_0) \\ G_2^0(ka, \Omega_0) \\ \vdots \\ G_N^{-N}(ka, \Omega_0) \\ \vdots \\ G_N^N(ka, \Omega_0) \end{bmatrix}, \tag{27}$$

which fulfills precisely the same role as (3). It is similarly possible to define a noise plus interference vector $\mathbf{N}(ka)$ in spherical harmonic space. Moreover, Yan et al. [40] demonstrated that the covariance matrix for the spherically isotropic noise field in spherical harmonic space can be expressed as

$$\begin{aligned} \mathbf{\Gamma}(ka) = 4\,\pi\sigma_{\text{SI}}^2\, \text{diag}\{&|b_0(ka)|^2, -|b_1(ka)|^2, -|b_1(ka)|^2, \\ &- |b_1(ka)|^2, |b_2(ka)|^2, \cdots, (-1)^N |b_N(ka)|^2\}, \end{aligned} \tag{28}$$

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

With the changes described above, all of the relations developed in Section 2.3 can be applied; the key intuition is that the physical microphones have been replaced by the spherical harmonics which have the very attractive property of the orthonormality as indicated by (21). In particular, the weights of the delay-and-sum beamformer can be calculated as in (5), the MVDR weights as in (8), and the diagonally loaded MVDR weights as in (9); the spherical harmonics super-directive beamformer is obtained by replacing $\mathbf{\Sigma_N}$ in (9) with (28).

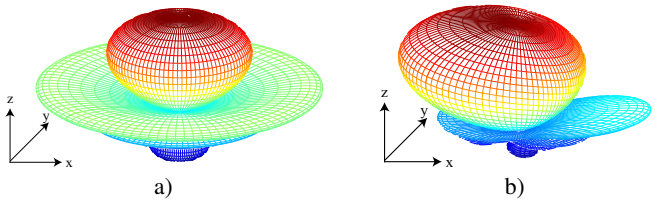


Fig. 13. Spherical MVDR beampatterns with a single plane wave interferer: a) radially symmetric, b) asymmetric.

4.2. Discretization

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

$$G_n^m(ka) = \sum_{s=0}^{S-1} \alpha_s G(ka, \Omega_s) \bar{Y}_n^m(\Omega_s), \tag{29}$$

$$G(ka, \Omega_s) = \sum_{n=0}^{N} \sum_{m=-n}^{n} G_n^m(ka) Y_n^m(\Omega_s), \tag{30}$$

where Ω_s indicates the position of microphone s and α_s is a quadrature constant. Typically N is limited such that $(N+1)^2 \leq S$ to prevent spatial aliasing [37].

Accordingly, the orthonormality condition (21) is approximated by the weighted summation, which causes orthonormality error [38]. In order to alleviate the error caused by discreteness, spatial sampling schemes [39] or beamformer's weights [38] must be carefully designed. In this article, we use a spherical microphone array with 32 equidistantly spaced sensors and set $\alpha_s = 4\pi/S$ in (29) for the experiments described later.

Shown in Figure 13 are two spherical three-dimensional MVDR beampatterns in the presence of a single plane wave interferer, one constrained to radial symmetry, and the other with no constraint. In both cases the look direction is $\Omega_0 = (0, 0)$ and the discrete interference is impinging from $\Omega_{\text{I}} = \left(\frac{\pi}{6}, \frac{-\pi}{4}\right)$. As is apparent from the figures, in order to place a null on the interferer, which is well inside the main lobe of the DS beamformer, it was necessary to allow large sidelobes.

4.3. Comparison of Linear and Spherical Arrays for DSR

As a spherical microphone array has—to the best of our knowledge—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 14 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 NaturalPoint's motion capture system, OptiTrack. For data capture we used an Eigenmike® which consists of 32 microphones embedded in a rigid spherical baffle of radius 4.2 cm; for further details see the website of mh acoustics, http://www.mhacoustics.com. Each sensor of the Eigenmike® is centered on the face of a truncated icosahedron. As a reference, the speech was also captured with the 64-channel linear microphone array described in Section 3.5. The aperture length of the linear array is 126 cm. The TIMIT data were used as test material. The test set consisted of 3,241 words uttered by 37 speakers for each recording position. The sampling rate of the data was 44.1 kHz. In the recording room, the reverberation time T_{60} was approximately 525 ms.

We used the same speech recognizer and decoding passes described in Section 3.4 for the experiments presented here. Tables 4 shows word error rates (WERs) for each beamforming algorithm in the case that the incident angles of the target signal to the array are 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 Tables 4 that every beamforming algorithm can provide the better recognition performance than the SAC after the adapted passes. It is also clear from the tables that super-directive 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 where the speaker position is nearly in front of the array, super-directive 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 in the spherical array. In the other configuration, however, spherical harmonics super-directive beamforming (Spherical SD BF) provided better results than the linear array because they can 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 the poor directivity.

5. CONCLUSIONS AND FUTURE DIRECTIONS

This contribution provided a comprehensive overview of representative microphone array methods for DSR. The article also presented

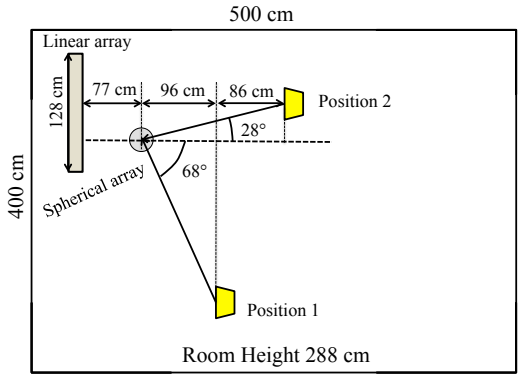


Fig. 14. Layout of the recording room.

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

a) Incident angle to the array is 28^{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	57.6	22.7	14.9	13.5
Spherical SD BF	44.8	15.5	11.3	9.7
Clean data	16.7	7.5	6.4	5.4

b) Incident angle to the array is 68^{o}.

Table 4. WERs for each beamforming algorithm.

recent progress in adaptive beamforming. The undesired effects such as signal cancellation and distortion of the target speech can be avoided by incorporating the fact that the distribution of speech signals is non-Gaussian into the framework of generalized sidelobe canceller beamforming. It was demonstrated that the state-of-the art DSR system can achieve recognition accuracy very comparable to that obtained by a close-talking microphone in a small vocabulary task. Finally, we discussed an emerging research topic, namely spherical microphone arrays. In terms of speech recognition performance, the spherical array was compared with the linear array. The results suggested that the compact spherical microphone array can achieve recognition performance comparable to a large linear array. In our view, a key research topic will be how we efficiently integrate different information sources such as faces in video data and turn-taking models into a DSR system.

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

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

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

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

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

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

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

[8] Heidi Christensen, Jon Barker, Ning Ma, and Phil Green, "The CHiME corpus: A resource and a challenge for computational hearing in multisource environments," in *Proc. Interspeech*, Makuhari, Japan, 2010.

[9] Tomohiro Nakatani, Takuya Yoshioka, Shoko Araki Marc Delcroix, and Masakiyo Fujimoto, "Logmax observation model

with MFCC-based spectral prior for reduction of highly nonstationary ambient noise," in *ICASSP 2012*, Kyoto, Japan, 2012.

[10] Felix Weninger, Martin Wöllmer, Jürgen Geiger, Björn Schuller, Jort Fe Gemmeke, Antti Hurmalainen, Tuomas Virtanen, and Gerhard Rigoll, "Non-negative matrix factorization for highly noise-robust asr: To enhance or to recognize?," in *ICASSP 2012*, Kyoto, Japan, 2012.

[11] Ramón Fernandez Astudillo, Alberto Abad, and Joao Paulo da Silva Neto, "Integration of beamforming and automatic speech recognition through propagationof the wiener posterior," in *ICASSP 2012*, Kyoto, Japan, 2012.

[12] Iain McCowan, Ivan Himawan, and Mike Lincoln, "A microphone array beamforming approach to blind speech separation," in *Proc. MLMI*, 2007.

[13] Shankar T. Shivappa, Bhaskar D. Rao, and Mohan M. Trivedi, "Audio-visual fusion and tracking with multilevel iterative decoding: Framework and experimental evaluation," *J. Sel. Topics Signal Processing*, vol. 4, no. 5, pp. 882–894, 2010.

[14] Matthias Wölfel, Kai Nickel, and John W. McDonough, "Microphone array driven speech recognition: Influence of localization on the word error rate," in *Proc. MLMI*, 2005, pp. 320–331.

[15] Ivan Jelev Tashev, *Sound Capture and Processing: Practical Approaches*, Wiley, Chichester, UK, 2009.

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

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

[18] Jingdong Chen, Jacob Benesty, and Yiteng Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP J. Adv. Sig. Proc.*, 2006.

[19] A. Brutti, M. Omologo, and P. Svaizer, "Comparison between different sound source localization techniques based on a real data collection," in *Proc. HSCMA*, Trento, Italy, 2008.

[20] Ulrich Klee, Tobias Gehrig, and John McDonough, "Kalman filters for time delay of arrival–based source localization," *EURASIP J. Adv. Sig. Proc.*, 2006.

[21] Tobias Gehrig, Kai Nickel, Hazim K. Ekenel, Ulrich Klee, and John McDonough, "Kalman filters for audio–video source localization," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2005.

[22] Tobias Gehrig, Ulrich Klee, John McDonough, Shajith Ikbal, Matthias Wölfel, and Christian Fügen, "Tracking and beamforming for multiple simultaneous speakers with probabilistic data association filters," in *Proc. Interspeech*, 2006.

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

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

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

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

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

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

[29] Kenichi Kumatani, John McDonough, Stefan Schacht, Dietrich Klakow, Philip N. Garner, and Weifeng Li, "Filter bank design based on minimization of individual aliasing terms for minimum mutual information subband adaptive beamforming," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, Nevada, U.S.A, 2008.

[30] Aapo Hyvärinen, Juha Karhunen, , and Erkki Oja, "Independent component analysis," *Wiley Inter-Science*, 2001.

[31] Kenichi Kumatani, John McDonough, Barbara Rauch, and Dietrich Klakow, "Maximum negentropy beamforming using complex generalized Gaussian distribution model," in *Proc. ASILOMAR*, Pacific Grove, CA, 2010.

[32] Kenichi Kumatani, John McDonough, Barbara Rauch, Philip N. Garner, Weifeng Li, and John Dines, "Maximum kurtosis beamforming with the generalized sidelobe canceller," in *Proc. Interspeech*, Brisbane, Australia, September 2008.

[33] Kenichi Kumatani, John McDonough, and Bhiksha Raj, "Maximum kurtosis beamforming with a subspace filter for distant speech recognition," in *Proc. ASRU*, 2011.

[34] Kenichi Kumatani, John McDonough, Jill Lehman, and Bhiksha Raj, "Channel selection based on multichannel cross-correlation coefficients for distant speech recognition," in *Proc. HSCMA*, Edinburgh, UK, 2011.

[35] Ernst Warsitz, Alexander Krueger, and Reinhold Haeb-Umbach, "Speech enhancement with a new generalized eigenvector blocking matrix for application in a generalized sidelobe canceller," in *Proc. ICASSP*, Las Vegas, NV, U.S.A, 2008.

[36] Kenichi Kumatani, Liang Lu, John McDonough, Arnab Ghoshal, and Dietrich Klakow, "Maximum negentropy beamforming with superdirectivity," in *European Signal Processing Conference (EUSIPCO)*, Aalborg, Denmark, 2010.

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

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

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

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

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

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

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