Microphone Array Processing for Distant Speech Recognition: Towards Real-World Deployment

Kenichi Kumatani*, Takayuki Arakawa†, Kazumasa Yamamoto‡, John McDonough§,
Bhiksha Raj¶, Rita Singh¶, and Ivan Tashev‖
* Disney Research, Pittsburgh, USA
E-mail: k_kumatani@ieee.org
† NEC Corporation, Kawasaki-shi, Japan
E-mail: t-arakawa@cp.jp.nec.com
‡ Toyohashi University of Technology, Toyohashi-shi, Japan
E-mail: kyama@tut.jp
§ Carnegie Mellon University, Voci Technologies, Inc., Pittsburgh, PA, USA
E-mail: johnmcd@cs.cmu.edu
¶ Carnegie Mellon Univerity, Pittsburgh, PA, USA
E-mail: {bhiksha,rsingh}@cs.cmu.edu
‖ Microsoft Research, Redmond, WA, USA
E-mail: ivantash@microsoft.com

***Abstract*—Distant speech recognition (DSR) holds out the promise of providing a natural human computer interface in that it enables verbal interactions with computers without the necessity of donning intrusive body- or head-mounted devices. Recognizing distant speech robustly, however, remains a challenge. This paper provides a 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 64-channel 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. Furthermore, we report the results of speech recognition experiments on data captured with a popular device, the Kinect [1]. Even for speakers at a distance of four meters from the Kinect, our DSR system achieved acceptable recognition performance on a large vocabulary task, a WER of 24.1%, beginning from a WER of 42.5% with a single array channel.**

I. Introduction

When the signals from the individual sensors of a microphone array with a known geometry are suitably combined, the array functions as a spatial filter capable of suppressing noise, reverberation, and competing speech. Such *beamforming* techniques have received a great deal of attention within the acoustic array processing community in the recent past [2], [3], [4], [5], [6], [7], [8].

Despite this effort, however, such techniques have often been ignored within the mainstream community working on distant speech recognition. As pointed out in [7], [8], [9], 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 ignores 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 [10], [11], [12], [13]. A notable exception was the PASCAL Speech Separation Challenge 2 [6], [14] which featured actual array recordings of real speakers. 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.

Until recently, an obstacle preventing the widespread use of microphone arrays in DSR applications was their prohibitive cost; this obstacle has been removed with the release of the Microsoft Kinect [1] platform, which provides a four-channel linear array for acoustic processing, together with an RGB camera and infrared depth sensor.

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.

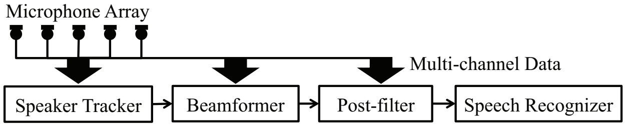

Fig. 1. Block diagram of a typical DSR system.

Towards this end, we present a series of empirical studies in DSR tasks with real speakers in real acoustic environments. The beamforming experiments are performed with a circular array with eight channels, a linear microphone array with sixty-four sensors and the Kinect with four microphones. In particular, results obtained with the Kinect are of great interest due to its reasonable price and capability of capturing multi-modal data including RGB and depth information.

This contribution complements Kumatani *et al* [9] with the new DSR experiments on the Kinect data. A more comprehensive review of microphone array processing for DSR will also appear in the book chapter McDonough and Kumatani [15].

The remainder of this article is organized as follows. In Section II, we provide an overview of a complete DSR system, which includes the fundamentals of array processing, speaker localization and conventional statistical beamforming techniques. In Section III, 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. In the final section of this work, we present our conclusions and plans for future work.

II. 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 [16].

A. 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 [17]; 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 [18, §5.5]. Finally, the placement of the sensors can deviate from their nominal positions. All of these factors degrade beamforming performance.

B. 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 [19, §8.2.1], localizing a source based on the application of high-resolution spectral estimation techniques such as subspace algorithms [20, §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 [21] and comparative studies on real data can be found in [22]. From the TDOA between the microphone pairs, the speaker's position can be computed using classical methods, namely, spherical intersection, spherical interpolation

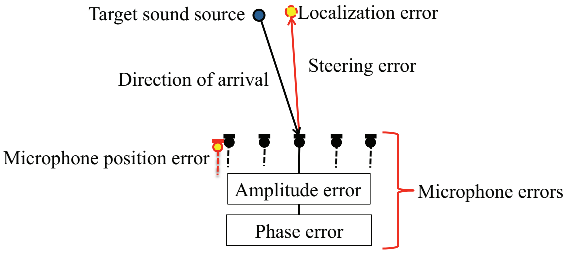

Fig. 2. Representative errors in microphone array processing.

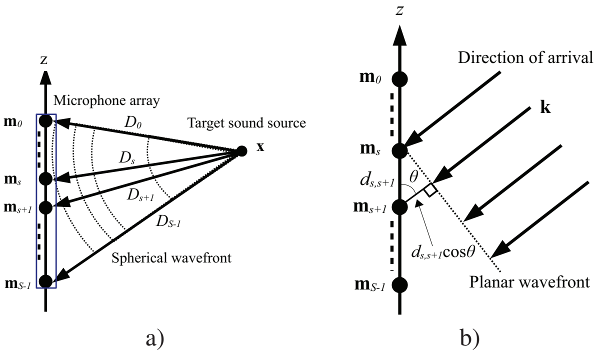

Fig. 3. Propagation of a) the spherical wave and b) plane wave.

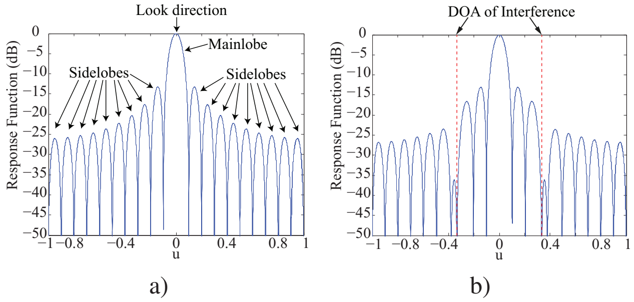

Fig. 4. Beampatterns of a) the delay-and-sum beamformer and b) MVDR beamformer as a function of $u = \cos\theta$ for the linear array; the noise covariance matrix of the MVDR beamformer is computed with the interference plane waves propagating from $u = \pm 1/3$.

or linear intersection [3, §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 [19, §10]. Klee et al. [23] demonstrated, 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 [24], and to track multiple simultaneous speakers [25].

C. Conventional Beamforming Techniques

Figure 3 shows how spherical and plane waves propagate in space. 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 assume that a plane wave with angular frequency ω is propagating and 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 $f(t)$ at time t, carried on a plane wave, reaches all sensors in an array, but not at the same time. Hence, let us form the vector

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

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

$$\mathbf{v}(\mathbf{k},\omega) \triangleq \begin{bmatrix} e^{-i\omega\tau_0} & e^{-i\omega\tau_1} & \cdots & e^{-i\omega\tau_{S-1}} \end{bmatrix}^T \qquad (3)$$

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

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

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

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

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

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

i.e., the output of the array is equivalent to the original signal in the absence of any interference or distortion in non-reverberant and no-noise environment. In general, this will be true for any weight vector achieving

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

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 $\theta \in [-\pi/2, +\pi/2]$ 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 [20, §7]. The weight vectors obtained under these conditions correspond to the *minimum variance distortionless response* (MVDR) beamformer, which has the well-known solution [3, §13.3.1] [26]

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

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 [3, §13.3.7]

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

which is referred to as *diagonal loading* where σ^2_d 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^2_d; hence the latter is typically set either based on experimentation or through an iterative procedure. Increasing σ^2_d decreases $\|\mathbf{w}_\text{DL}\|$, which implies that the *white noise gain* (WNG) also increases [27]; 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 [19, §4]. The MVDR beamformer with the diffuse noise model is called the *super-directive beamformer* or *time invariant beamformer with cylindrical noise* [3, §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

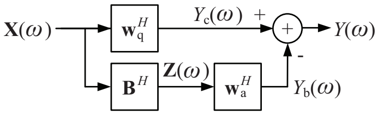

Fig. 5. Generalized sidelobe cancellation beamformer in the frequency domain.

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

where $d_{m,n}$ is the distance between the mth and nth elements of the array, and 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

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

where σ^2_SI 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 [20, §6.2.3]. Secondly, and more importantly, by placing a Wiener filter [28, §2.2] on the output of the MVDR beamformer, the *minimum mean-square error* (MMSE) beamformer is obtained [20, §6.2.2]. Such *post-filters* are important because it has been shown that they can yield significant reductions in error rate [6], [29], [30], [31]. Of the several post-filtering methods proposed in the literature [32], the *Zelinski post-filtering* [33], [34] 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 Section III-D.

The MVDR beamformer can be implemented in generalized sidelobe canceller (GSC) configuration [20, §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

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

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) [3, §13.6]. The blocking matrix $\mathbf{B}$ is chosen to be orthogonal to $\mathbf{w}_\text{q}$, such that $\mathbf{B}^H\,\mathbf{w}_\text{q} = \mathbf{0}$. This orthogonality implies that the distortionless constraint will be satisfied for any choice of $\mathbf{w}_\text{a}$.

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 [35].

For the empirical studies reported in Section III-D, subband analysis and synthesis were performed with a uniform DFT filter bank based on the modulation of a single prototype impulse response [3, §11][36]. 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 [36] 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

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

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.

III. 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 [35]; it has been demonstrated that such HOS beamformers do *not* suffer from signal cancellation.

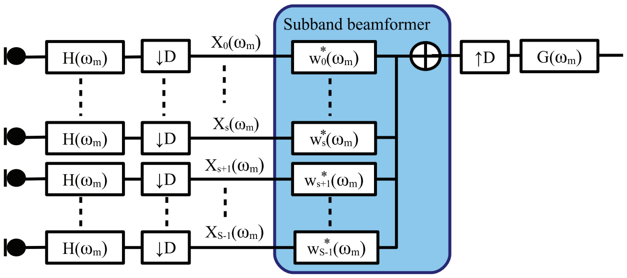

Fig. 6. Schematic view for beamforming in the subband domain.

In this section, we introduce a concept of non-Gaussianity and describe how the fine 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 [37]; thereafter we discuss beamforming algorithms based on maximizing non-Gaussian optimization criteria.

A. Motivation for Maximizing non-Gaussianity

The *central limit theorem* [38] 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* [38, §8].

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

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

where β_K is a positive constant, which is typically set to $\beta_\text{K} = 3$ 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* [38]. The differential entropy for continuous complex-valued RVs Y with the pdf $p_Y(\cdot)$ is defined as

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

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

$$\text{neg}(Y) \triangleq H(Y_\text{gauss}) - H(Y), \tag{15}$$

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

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

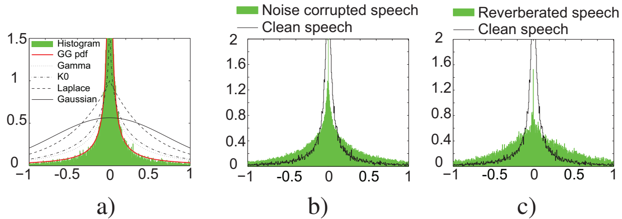

Fig. 7. Histograms of real parts of subband frequency components of clean speech and a) pdfs, b) noise-corrupted speech and c) reverberated speech.

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 $H(Y)$ 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 III-D. The form of the GG-pdf and entropy is described in [39]. 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 [38, §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.

B. 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

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

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 [35], [39] for maximum negentropy (MN) beamforming and [40] for maximum kurtosis (MK) beamforming.

As shown in [35], 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.

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.

C. 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 [20, §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 [41].

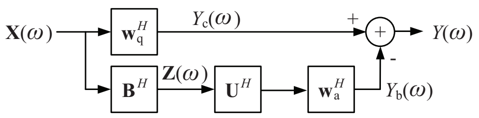

Fig. 8. Maximum kurtosis beamformer with the subspace filter.

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

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

The difference between (12) and (18) is the subspace filter between the blocking matrix and active weight vector. The

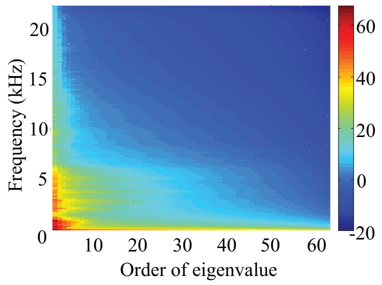

Fig. 9. Eigenvalues sorted in descending order over frequencies.

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 [20, §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 [42]. 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, 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.

D. *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 [5] for details of the data collection apparatus. The size of the recording room was 650 × 490 × 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 [25]. 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 [35]. 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) [3, §9] and *constrained maximum likelihood linear regression* (CMLLR) parameters [3, §9] for each speaker, then redecode with the conventional ML acoustic model; (3) estimate VTLN, CMLLR and *maximum likelihood linear regression* (MLLR) [3, §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 [3, §8.1]. The standard WSJ trigram language model was used in all passes of decoding.

Table I 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 I 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

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 I
WORD ERROR RATES FOR EACH BEAMFORMING ALGORITHM AFTER EVERY DECODING PASS.

recognition performance than the SOS-based beamformers, such as the super-directive beamformer (SD BF) [3, §13.3.4], the MVDR beamformer (MVDR BF) and the generalized eigenvector beamformer (GEV BF) [43]. 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 III-B, 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 [44]. We observe from Table I that the maximum negentropy beamformer with super-directive beamformer (SD MN BF) provided the best recognition performance in this task.

E. *Effect of HOS-based Beamforming with Subspace Filtering*

In this section, we investigate effects of MK beamforming with subspace filtering. The *Copycat* data [42] were used as test material here. In the Copycat, children repeat what an adult instructor said without a transcription. The speech material in this corpus was captured with the 64-channel linear microphone array; the sensors were arranged equidistantly 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 [41]. The decoder used here consists of three passes; the first and second passes are the same as the ones described in Section III-D but the third pass includes processing of the third and fourth passes described in Section III-D.

	Pass (%WER)					
	1		2		3	
Algorithm	Adult	Child	Adult	Child	Adult	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 II
WORD ERROR RATES (WERS) FOR EACH DECODING PASS.

Algorithm	Block size (second)	Pass (%WER)			
		2		3	
		Adult	Child	Adult	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 III
WERS AS A FUNCTION OF AMOUNTS OF ADAPTATION DATA.

Table II 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 (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 II that the maximum kurtosis beamformer with subspace filtering achieved the best recognition performance.

Table III 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 III 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 III 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 III-C. Note that the solution of the eigendecomposition does not depend on the initial value in contrast to the gradient-based numerical optimization algorithm.

F. *Beamforming with the Kinect sensor*

We used the Kinect sensor [1] in order to confirm the effectiveness of our beamforming methods. The Kinect has the non-uniform linear array with four microphones, the RGB camera and IR depth sensor. The Kinect Software Development Kit (KSDK) version 1.5 that implements speaker localization and beamforming is also available.

Figure 10 and 11 show recording configuration and a picture of the recording system for the data collection. One Kinect is

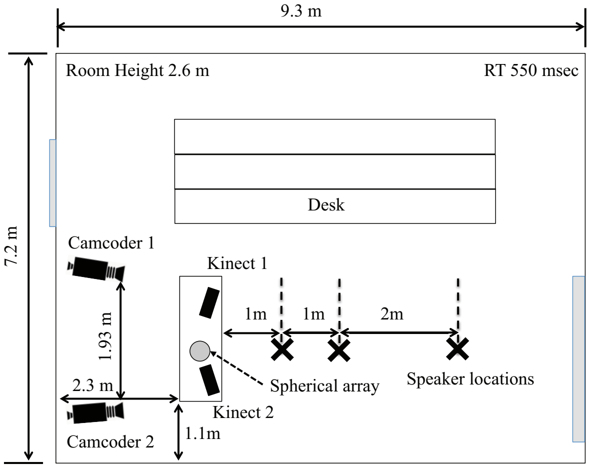

Fig. 10. Apparatuses for the Kinect recording.

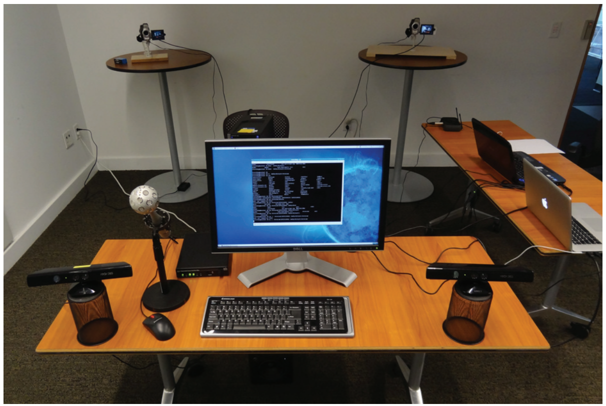

Fig. 11. Recording System.

used for capturing the four-channel data and the other is put for recording the output of the Kinect beamformer in real-time. We used a default setting of the Kinect beamformer provided in the KSDK. In recording sessions, eleven human subjects are asked to read 25 sentences of the Wall Street Journal corpus at two different positions in order to investigate the sensitivity of recognition performance against a distance between the array and speaker. The close talking microphone data were also recorded as a reference. As shown in Fig. 10, the distances are 1, 2, or 4 meters. The test data consisted of 6,948 words in total. The reverberation time T_{60} was approximately 550 milliseconds. No noise was artificially added to the captured data, as natural noise from air conditioners, computers and other speakers was already present.

After the speakers positions were obtained with the speaker tracking system [23], we performed the speaker beamforming algorithms on the pre-recorded Kinect data. The beamformed signal was further enhanced with Zelinski post-filtering. We then ran the multiple-pass speech recognizer described in III-D on the enhanced speech data. Table IV shows word error rates obtained with each beamforming algorithm as a function of distances between the Kinect and speaker. As references, the WERs obtained with the single array channel (SAC), default Kinect beamformer and close-talking microphone (CTM) are shown in Table IV. It is clear from Table IV that every beamforming algorithm can provide better recognition performance than the SAC. Especially in the case that the distance between the array and speaker is 4 meters, the improvement by beamforming is significant. It is also clear from Table IV that the super-directive beamformer with the active weight vector adjusted based on the maximum kurtosis criterion (SD-MK BF) achieves the best recognition performance. Note that we did not tune any parameters for the KSDK's beamformer.

Algorithm	Distance	Pass (%WER)			
		1	2	3	4
Single array channel (SAC)	1m	72.3	34.3	26.8	24.7
	2m	82.2	49.6	35.8	32.4
	4m	90.8	68.6	49.3	42.5
Delay-and-sum BF (D&S BF)	1m	43.7	27.1	23.1	21.8
	2m	55.7	34.2	27.8	25.2
	4m	72.3	41.8	29.0	26.7
Super-directive BF (SD BF)	1m	42.2	28.8	23.5	22.0
	2m	50.8	31.6	26.0	24.5
	4m	66.2	39.2	28.1	25.1
Super-directive maximum kurtosis BF (SD-MK BF)	1m	38.6	25.4	21.7	20.0
	2m	49.2	30.4	24.8	23.5
	4m	65.8	36.7	27.4	24.1
Default Kinect BF	1m	48.3	28.6	23.7	21.3
	2m	71.3	41.5	33.2	30.3
	4m	80.5	56.1	41.2	37.4
CTM	Avg.	31.7	20.9	16.4	15.6

TABLE IV
WERS AS A FUNCTION OF DISTANCES BETWEEN THE SPEAKERS AND KINECT.

In the large vocabulary continuous speech recognition task, the distant speech recognizer with beamforming still lags behind the close talking microphone. However, the recognition performance can be acceptable in applications that do not require recognizing every word precisely such as dialogue systems.

IV. Conclusions and Future Directions

This paper provided a comprehensive overview of representative microphone array methods for DSR. The article also presented 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 conducted speech recognition experiments on data collected with the Kinect [1]. The results suggested that the practical recognition performance can be obtained with the reasonably priced device.

References

[1] "Microsoft Kinect for Windows," http://www.microsoft.com/en-us/kinectforwindows/.

[2] 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.

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

[4] 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.

[5] M. Lincoln, I. McCowan, I. Vepa, and H. K. Maganti, “The multichannel 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.

[6] 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.

[7] 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.

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

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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

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

[17] 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.

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

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

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

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

[22] 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.

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

[24] 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.

[25] 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.

[26] O. Frost III, “An algorithm for linearly constrained adaptive array processing,” *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–934, August 1972.

[27] 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.

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

[29] 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.

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

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

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

[33] Rainer Zelinski, “A microphone array with adaptive post-filtering for noise reduction in reverberant rooms,” in *Proc. ICASSP*, 1988, pp. 2578–2581.

[34] 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.

[35] 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.

[36] 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.

[37] Ivan Tashev and Alex Acero, “Statistical modeling of the speech signal,” in *Proc. International Workshop on Acoustic Echo and Noise Control*, Tel Aviv, Israel, 2010.

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

[39] 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.

[40] 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.

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

[42] 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.

[43] 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.

[44] 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.