Beamforming with a Maximum Negentropy Criterion

Kenichi Kumatani, John McDonough, Barbara Rauch, Dietrich Klakow,
Philip N. Garner and Weifeng Li

***Abstract*— In this paper, we address a beamforming application based on the capture of far-field speech data from a single speaker in a real meeting room. After the position of the speaker is estimated by a speaker tracking system, we construct a subband-domain beamformer in *generalized sidelobe canceller* (GSC) configuration. In contrast to conventional practice, we then optimize the *active weight vectors* of the GSC so as to obtain an output signal with *maximum negentropy* (MN). This implies the beamformer output should be as non-Gaussian as possible. For calculating negentropy, we consider the Γ and the generalized Gaussian (GG) pdfs. After MN beamforming, Zelinski post-filtering is performed to further enhance the speech by removing residual noise. Our beamforming algorithm can suppress noise and reverberation without the signal cancellation problems encountered in the conventional beamforming algorithms. We demonstrate this fact through a set of acoustic simulations. Moreover, we show the effectiveness of our proposed technique through a series of far-field automatic speech recognition experiments on the *Multi-Channel Wall Street Journal Audio Visual Corpus* (MC-WSJ-AV), a corpus of data captured with real far-field sensors, in a realistic acoustic environment, and spoken by real speakers. On the MC-WSJ-AV evaluation data, the delay-and-sum beamformer with post-filtering achieved a word error rate (WER) of 16.5%. MN beamforming with the Γ pdf achieved a 15.8% WER, which was further reduced to 13.2% with the GG pdf, whereas the simple delay-and-sum beamformer provided a WER of 17.8%. To the best of our knowledge, no lower error rates at present have been reported in the literature on this ASR task.**

***Index Terms*— microphone arrays, beamforming, speech recognition, speech enhancement, source separation**

I. Introduction

There has been great and growing interest in microphone array processing for hands-free speech recognition [1], [2], [3]. Such techniques have the potential to relieve users from the necessity of donning close talking microphones (CTMs) before dictating or otherwise interacting with automatic speech recognition (ASR) systems. Beamforming is a promising technique for far-field speech recognition. A conventional beamformer in *generalized sidelobe canceller* (GSC) configuration is structured such that the direct signal from a desired direction is undistorted [4, §13.3.7]. Typical GSC beamformers consist of three blocks, a *quiescent vector*, *blocking matrix* and *active weight vector*. The quiescent vector is calculated to provide unity gain for the direction of interest. The blocking matrix is usually constructed in order to keep a distortionless constraint for the signal filtered with the quiescent vector. Subject to this constraint, the total output power of the beamformer is minimized through the adjustment of an active weight vector, which effectively places a null on any source of interference, but can also lead to undesirable *signal cancellation* [5]. To avoid the latter, many algorithms have been developed. These approaches fall into one of the following categories:

- updating the active weight vector only when noise signals are dominant [6], [7], [8];
- constraining the update formula for the active weight vector with the leaky least mean square (LMS) algorithm [9], [10] or with power of outputs of the blocking matrix [11];
- using multi-channel target signals received by the microphone array and correlation matrices of the clean and noise corrupted target signals in a calibration phase, [12];
- blocking the leakage of desired signal components into the sidelobe canceller by appropriately designing the blocking matrix [11], [13], [14], [15];
- taking speech distortion due to the leakage of a target signal into account using a multi-channel Wiener filter which aims at minimizing a weighted sum of residual noise and speech distortion terms [16]; and
- using acoustic transfer functions from a desired source to microphones instead of merely compensating for the time delays [8], [15], [17], [18].

Low et al. [19] proposed an approach that differs from the traditional GSC beamforming algorithms. They combined a blind source separation (BSS) technique [20] and an adaptive noise canceller with the modified leaky LMS algorithm. Their algorithm first estimates the unmixing matrix with the information maximization technique [20]. The permutation problem is alleviated through use of the directivity pattern [21], and the scaling ambiguity is eliminated by forcing the determinant of the unmixing matrices to unity [22]. The output channel with the highest kurtosis value is then taken as the target speech and the others are labeled as reference signals. The adaptive noise canceller finally removes any components that are correlated to the reference signals, which also leads to the signal cancellation problem. To prevent the latter, Low

This work was supported by the European Union (EU) under the integrated projects AMIDA, *Augmented Multi-party Interaction with Distance Access*, contract number IST-033812, and by the Federal Republic of Germany under the research training network IRTG 715 “Language Technology and Cognitive Systems”, funded by the German Research Foundation (DFG). Kenichi Kumatani is with the Institute for Computer Science and Engineering, Intelligent Sensor-Actuator Systems (ISAS) at the University of Karlsruhe in Karlsruhe, Germany and with the IDIAP Research Institute in Martigny, Switzerland. John McDonough is with ISAS at the University of Karlsruhe and with Spoken Language Systems at Saarland University in Saarbrücken, Germany. Barbara Rauch and Dietrich Klakow are with Spoken Language Systems at Saarland University in Saarbrücken, Germany. Philip Garner and Weifeng Li are with the IDIAP Research Institute.

et al. proposed the modified leaky LMS algorithm, which adjusts the step-size used for the weight update with a non-linear function. In their algorithm, the weights of the unmixing matrix for extracting the desired signal can be regarded as the block of the upper branch in the GSC structure and the other weights can be associated with the blocking matrix. Then, the active noise canceller corresponds to the active weight vector. Therefore the method proposed by Low et al. could be viewed as a GSC beamforming algorithm without the distortionless constraint. The BSS algorithms only provide a local solution, however, which is highly dependent on the initial values. Moreover, a lower bound on the performance of the speech enhancement cannot be given. The unmixing matrix obtained with this technique may fail to extract the target signal in some situations. Moreover, it could happen that permutation and scaling ambiguity problems are still present after the weights have converged. Such uncertain behavior would be unacceptable for many applications.

Parra and Alvino [23] proposed the *geometric source separation* (GSS) algorithm for the source separation problem. The GSS algorithm estimates the unmixing matrix with the geometric constraint which can implicitly solve the permutation and scaling ambiguity problems. The current authors noted in [2] that this algorithm is equivalent to constructing two GSC beamformers and estimating the active weight vectors so as to decorrelate the outputs of the two beamformers. The current authors also proposed a beamforming algorithm whereby the active weight vectors of two beamformers are adjusted in order to achieve minimum mutual information (MMI) between the outputs of the beamformers [2]. The mutual information criterion was noted to yield a optimization criterion *similar* to that used in the GSS algorithm under a Gaussian assumption. One of the principal advantages of the MMI formulation is that it can be readily extended to non-Gaussian pdfs.

In this work, we consider *negentropy* as a criterion for estimating the active weight vectors in a GSC. Negentropy indicates how far a probability density function (pdf) of a particular signal is from Gaussian. The pdf of speech is in fact super-Gaussian [2], [24], [25], but it becomes closer to Gaussian when the speech is corrupted by noise or reverberation. Hence, in adjusting the active weight vector of the GSC to provide a signal with the highest possible negentropy, we hope to remove or suppress noise and reverberation. As we will demonstrate, the *maximum negentropy* (MN) beamformer can achieve this goal without the signal cancellation problem encountered in conventional beamforming algorithms [5]. Moreover, our technique can circumvent the permutation and scaling ambiguity problems by maintaining a distortionless constraint in the look direction. For calculating negentropy, we consider the Γ and generalized Gaussian (GG) pdfs, and investigate the suitability of each for this task. After MN beamforming, *Zelinski* post-filtering is performed to further enhance the speech by removing residual noise [26]. The Zelinski post-filtering technique is efficient for removing incoherent noise since it assumes zero-correlation between the noise on different sensors. It should be noted, however, that such an assumption may be inappropriate in several applications [27], [28], [29].

We demonstrate the effectiveness of our proposed technique through a series of far-field automatic speech recognition experiments on the *Multi-Channel Wall Street Journal Audio Visual Corpus* (MC-WSJ-AV) collected under the European Union integrated project *Augmented Multi-party Interaction* (AMI) [1]. The data was recorded in a real meeting room, and hence contains noise from computers, fans, and other apparatus in the room. Moreover, some recordings include noise coming from outside the meeting room, such as that produced by passing cars or speakers in an adjacent room. The test data is neither artificially convolved with measured impulse responses nor unrealistically mixed with separately-recorded noise.

The balance of this work is organized as follows. We describe the super-Gaussian pdfs which are used for calculating the negentropy in Section II. In particular, Section II shows that the distribution of clean speech is not Gaussian but super-Gaussian and the pdf of noise-corrupted speech becomes closer to Gaussian. Section III reviews the definition of entropy and negentropy. Section IV illustrates the speech distribution modeled with the GG pdf. In Section V, we discuss the maximum negentropy beamforming criterion and then derive the gradient relations required for optimizing the active weight vector of the GSC. In Section VI, we demonstrate that the proposed beamforming algorithm has no signal cancellation problem through a set of acoustic simulations. In Section VII, we describe the results of far-field automatic speech recognition experiments. Finally, in Section VIII, we present our conclusions and plans for future work.

II. Modeling Subband Samples of Speech with Super-Gaussian Probability Density Functions

Here we review theoretical arguments and empirical evidence that subband samples of speech, like nearly all other information bearing signals, are *not* Gaussian-distributed [30].

The entire field of *independent component analysis* (ICA) is founded on the assumption that all signals of real interest are *not* Gaussian-distributed [30]. Briefly, the reasoning is grounded on two points:

1. The *central limit theorem* states that the pdf of the sum of independent random variables (r.v.s) will approach Gaussian in the limit as more and more components are added, *regardless* of the pdfs of the individual components. This implies that the sum of several r.v.s will be closer to Gaussian than any of the individual components. Thus, if the original independent components comprising the sum are sought, one must look for components with pdfs that are the *least* Gaussian.
2. The *entropy* for a continuous complex-valued r.v. Y, is defined as
$$H(Y) \triangleq -\int p_Y(v) \log p_Y(v) dv = -\mathcal{E}\left\{\log p_Y(v)\right\}, \tag{1}$$
where $p_Y(.)$ is the pdf of Y. Entropy is the basic measure of information in *information theory* [31]. It is well known that a Gaussian r.v. has the highest entropy of all r.v.s with a given variance [31, Thm. 7.4.1], which also holds for complex Gaussian r.v.s [32, Thm. 2]. Hence, a Gaussian r.v. is,

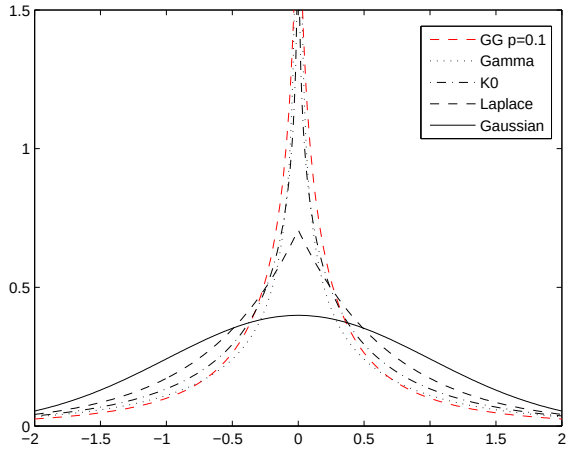


Fig. 1. Gaussian and super-Gaussian pdfs.

in some sense, the least *predictable* of all r.v.s. Information-bearing signals, on the other hand, are redundant and thus contain structure that makes them more predictable than Gaussian r.v.s. Hence, if an information-bearing signal is sought, one must once more look for a signal that is *not* Gaussian.

The fact that the pdf of speech is super-Gaussian has often been reported in the literature [2], [24], [25]. Noise, on the other hand, is more nearly Gaussian-distributed. In fact, the pdf of the sum of several super-Gaussian r.v.s. becomes closer to Gaussian. Thus, a mixture consisting of a desired signal and several interfering signals can be expected to be nearly Gaussian-distributed.

The Gaussian and four super-Gaussian univariate pdfs are plotted in Fig. 1. From the figure, it is clear that the Laplace, K_0, Γ, and GG densities exhibit the "spikey" and "heavy-tailed" characteristics that are typical of super-Gaussian pdfs. This implies that they have a sharp concentration of probability mass at the mean, relatively little probability mass as compared with the Gaussian at intermediate values of the argument, and a relatively large amount of probability mass in the tail; i.e., far from the mean.

Fig. 2 shows a histogram of the real parts of subband samples of speech at f_s = 800 Hz. To generate these histograms, we used 43.9 minutes of clean speech recorded with a close-talking microphone (CTM) from the development set of the Speech Separation Challenge, Part 2 (SSC2) [1]. Shown in Fig. 2 are also plots of the Gaussian, Laplace, K_0, Γ, and generalized Gaussian pdfs. For this plot, the shape parameter of the GG pdf was estimated from training data. It is clear from Fig. 2 that the distribution of clean speech is not Gaussian but super-Gaussian. Fig. 2 also suggests that the GG pdf can be suitable for modeling subband samples of speech.

Fig. 3 shows the histogram of magnitude in the subband domain[1]. We can see from Fig. 3 that the GG pdf can model the distribution of magnitude in the subband domain very well.

Fig. 4 shows histograms of real parts of subband components calculated from clean speech and noise-corrupted speech. It is clear from this figure that the pdf of the noise-corrupted speech has less probability mass around the center spike, and less probability mass in the tail than the clean speech, but more probability mass in intermediate regions. This indicates that the pdf of the noise-corrupted signal, which is in fact the sum of the speech and noise signals, is closer to Gaussian than that of clean speech. Fig. 5 shows histograms of clean speech and reverberant speech in the subband domain. In order to produce the reverberant speech, a clean speech signal was convolved with an impulse response measured in a room; see Lincoln *et al.* [1] for the configuration of the room. We can observe from Fig. 5 that the pdf of reverberated speech is also closer to Gaussian than the original clean speech.

We also present a histogram of magnitude of noise corrupted speech in Fig. 6 and that of reverberant speech in Fig. 7. We can again see from Fig. 6 and Fig. 7 that the pdfs of corrupted speech have less probability mass around the mean and less probability mass in the tail, but once more more probability mass in intermediate regions. Interestingly, Fig. 7 shows that the peak of the histogram of the speech is shifted from zero to the right by the reverberation effect.

These facts would indeed support the hypothesis that seeking an enhanced speech signal that is maximally non-Gaussian is an effective way to suppress the distorting effects of noise and reverberation.

A. Super-Gaussian pdf derived from the Meijer G-function

As noted by Brehm and Stammler [33], it is useful to model speech as a *spherically-invariant random process* (SIRP), because such processes are completely characterized by their first and second order moments. Moreover, Brehm and Stammler [33] noted that the Laplace, K_0, and Γ pdfs can all be represented as *Meijer G-functions*, which is useful for two reasons. Firstly, this implies that multivariate pdfs of all orders can be readily derived from the univariate pdf. Secondly, such variates can be extended to the case of complex r.v.s.

For the empirical studies reported here, a Γ pdf was used, as it achieved a higher likelihood than the other two named pdfs, namely, Laplace, and K_0 [2]. For the Γ pdf, the complex univariate pdf *cannot* be expressed in closed form in terms of elementary or even special functions. As explained in [2], however, it is possible to derive Taylor series expansions that enable the required variates to be calculated to arbitrary accuracy. Similarly, the differential entropy for the Γ pdf can also not be expressed in closed form. Hence, in order to use the Γ pdf, it is necessary to replace the exact differential entropy

[1]The pdfs in Fig. 3 are generally defined over the interval (-∞, +∞). Precisely speaking, the double-sided pdfs should be modified in order to model magnitude whose value is always positive. This is easily done by multiplying both sides by a factor of two and redefining the interval as [0, +∞). Such modifications, however, are not necessary in our algorithm in that the factor of two in the normalization is constant in the log-likelihood domain and has no effect on the gradient algorithm.

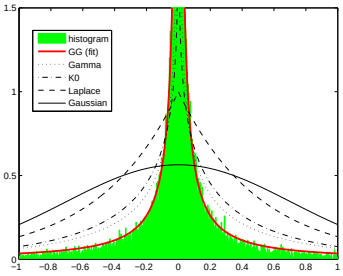


Fig. 2. Histogram of real parts of subband components and pdfs.

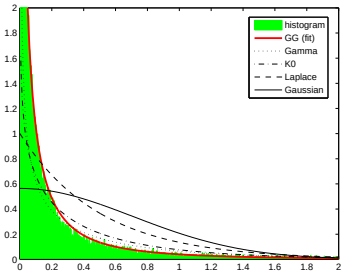


Fig. 3. Histogram of magnitude in the subband domain and pdfs.

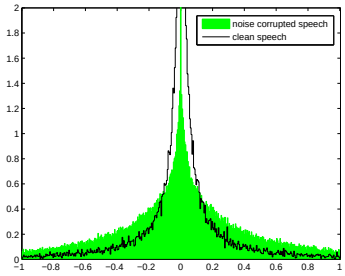


Fig. 4. Histograms of clean speech and noise corrupted speech in the subband domain.

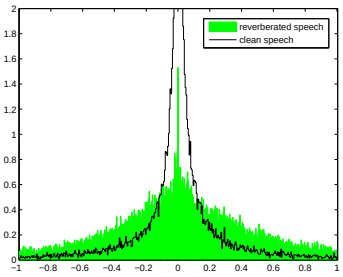


Fig. 5. Histograms of clean speech and reverberant speech in the subband domain.

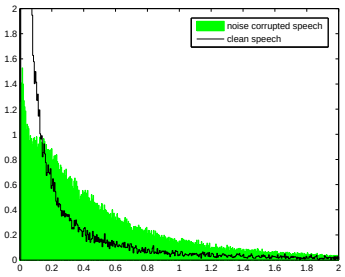


Fig. 6. Histograms of the magnitude of clean speech and noise corrupted speech in the subband domain.

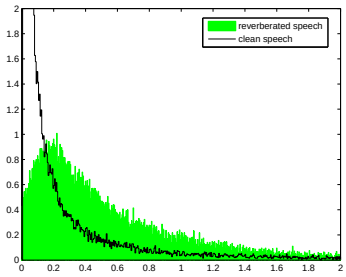


Fig. 7. Histograms of magnitude of clean speech and reverberated speech in the subband domain.

with the *empirical entropy*

$$H(Y) = -\mathcal{E}\left\{\log p_Y(v)\right\} \approx -\frac{1}{N}\sum_{n=0}^{N-1}\log p_Y(Y_n), \quad (2)$$

where Y_n is an observed subband sample.

B. Generalized Gaussian pdf

Due to its definition as a contour integral, finding maximum likelihood estimates for the parameters of a Meijer G-function must necessarily devolve to a grid search over the relevant parameter space [33]. Instead, it might be better to use a simple super-Gaussian pdf whose parameters can easily be adjusted so as to match the subband samples. The generalized Gaussian (GG) pdf is well-known and finds frequent application in the BSS and ICA fields. Moreover, it subsumes the Gaussian and Laplace pdfs as special cases. The GG pdf with zero mean for a real-valued r.v. y can be expressed as

$$p_{\mathrm{GG}}(y) = \frac{1}{2\Gamma(1+1/p)A(p,\hat{\sigma})}\exp\left[-\left|\frac{y}{A(p,\hat{\sigma})}\right|^p\right], \quad (3)$$

where p is the *shape parameter*, $\hat{\sigma}$ is the *scale parameter* which controls how fast the tail of the pdf decays, and

$$A(p,\hat{\sigma}) \;=\; \hat{\sigma}\left[\frac{\Gamma(1/p)}{\Gamma(3/p)}\right]^{1/2}. \quad (4)$$

In (4), $\Gamma(.)$ is the gamma function. Note that the GG with $p=1$ corresponds to the Laplace pdf, and that setting $p=2$ yields the Gaussian pdf, whereas in the case of $p\to+\infty$ the GG pdf converges to a uniform distribution.

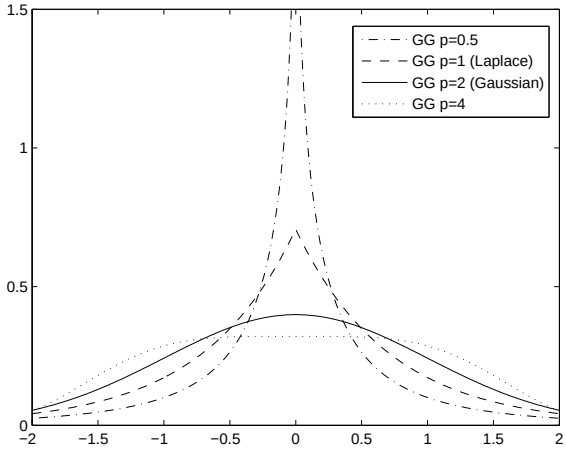


Fig. 8. The generalized Gaussian (GG) pdfs.

Fig. 8 shows the GG pdf with the same scale parameter $\hat{\sigma}^2=1$ and different shape parameters, $p=0.5,1,2,4$. From the figure, it is clear that a smaller shape parameter yields a spikier pdf with a heavier tail.

The differential entropy of the GG pdf for the real-valued r.v. y is obtained with the help of *Mathematica* [34] as

$$\begin{aligned} H_{\mathrm{GG}}(y) &= -\int_{-\infty}^{+\infty} p_{\mathrm{gg}}(\xi)\log p_{\mathrm{gg}}(\xi)d\xi \\ &= \frac{1}{p} + \log\left[2\Gamma(1+1/p)A(p,\hat{\sigma})\right]. \end{aligned} \quad (5)$$

Maximum likelihood (ML) estimates of the shape and scale parameters can be determined from a set of training data, as described in the next section.

C. Methods for Estimating Scale and Shape Parameters

Among several methods for estimating the shape parameter p of the GG pdf [35][36], the moment and ML methods are arguably the most straightforward. In this work, we used the moment method in order to initialize the parameters of the GG pdf and then updated them with the ML estimate [36]. The shape parameters are estimated from training samples offline and are then held fixed during beamforming. The shape parameters are estimated independently for each subband, as the optimal pdf is frequency-dependent.

For a set $\mathcal{Y} = \{y_0, y_1, \ldots, y_{N-1}\}$ of N real-valued training samples, the log-likelihood function under the GG pdf can be expressed as

$$\begin{aligned} l(\mathcal{Y};\hat{\sigma},p) = &-N\log\{2\Gamma(1+1/p)A(p,\hat{\sigma})\} \\ &-\frac{1}{A(p,\hat{\sigma})^p}\sum_{n=0}^{N-1}|y_n|^p. \end{aligned} \tag{6}$$

In this work, we considered three kinds of training sample y_n, namely, the magnitude as well as the real and imaginary parts of the subband samples of speech.

The parameters $\hat{\sigma}$ and p can be obtained by solving the following equations:

$$\frac{\partial l(\mathcal{Y};\hat{\sigma},p)}{\partial\hat{\sigma}} = -\frac{N}{\hat{\sigma}} + \frac{p}{\hat{\sigma}^{p+1}}\left[\frac{\Gamma(1/p)}{\Gamma(3/p)}\right]^{-\frac{p}{2}}\sum_{n=0}^{N-1}|y_n|^p = 0, \tag{7}$$

$$\begin{aligned} \frac{\partial l(\mathcal{Y};\hat{\sigma},p)}{\partial p} =& Na(p) - \sum_{n=0}^{N-1}\left(\frac{|y_n|}{A(p,\hat{\sigma})}\right)^p \\ &\times\left[\log\left\{\frac{|y_n|}{A(p,\hat{\sigma})}\right\} + b(p)\right] = 0, \end{aligned} \tag{8}$$

where

$$\begin{aligned} a(p) &= (p^{-2}/2)[2\Psi(1+1/p) + \Psi(1/p) - 3\Psi(3/p)], \\ b(p) &= (p^{-1}/2)[\Psi(1/p) - 3\Psi(3/p)], \end{aligned}$$

and $\Psi(.)$ is the digamma function. By solving (7) for $\hat{\sigma}$, we obtain

$$\hat{\sigma} = \left[\frac{\Gamma(3/p)}{\Gamma(1/p)}\right]^{1/2}\left(\frac{p}{N}\sum_{n=0}^{N-1}|y_n|^p\right)^{1/p}. \tag{9}$$

Due to the presence of the special functions, it is impossible to solve (8) for p explicitly. Varanasi [37] showed, however, that (8) has a unique root given the scale parameter. Hence, the gradient descent algorithm [38] can be used to find the unique solution which maximizes the likelihood. The solution of (8) can be also obtained with the secant algorithm [34], [37]. The estimation of the parameters is repeated until the log-likelihood function (6) converges.

III. Negentropy and Kurtosis

There are two popular criteria for measuring non-Gaussianity, namely, negentropy and kurtosis, both of which are frequently used in the field of ICA [30].

The negentropy of a complex-valued r.v. Y is defined as

$$J(Y) \triangleq H(Y_{\text{gauss}}) - H(Y) \tag{10}$$

where Y_{gauss} is a Gaussian variable which has the same variance σ_Y^2 as Y. The entropy of Y_{gauss} can be expressed as

$$H(Y_{\text{gauss}}) = \log\left|\sigma_Y^2\right| + 2\left(1+\log 2\pi\right). \tag{11}$$

In Section II, we calculated $H(Y)$ in (10) with two super-Gaussian distributions, namely, the Γ and GG pdfs. Note that negentropy is non-negative, and zero if and only if Y has a Gaussian distribution.

The *excess kurtosis* or simply kurtosis of a complex-valued r.v. Y with zero mean is defined as

$$\text{kurt}(Y) \triangleq \mathcal{E}\{|Y|^4\} - 3(\mathcal{E}\{|Y|^2\})^2.$$

The Gaussian pdf has zero kurtosis, pdfs with positive kurtosis are super-Gaussian, those with negative kurtosis are *sub-Gaussian*. Of the three super-Gaussian pdfs in Fig. 1, the Γ pdf has the highest kurtosis, followed by the K_0, then by the Laplace pdf. As is clear from Fig. 1, as the kurtosis increases, the pdf becomes more spikey and heavy-tailed. Note that the kurtosis of the GG pdf can be controlled by adjusting the shape parameter p, as explained in Section IV.

Kurtosis can be calculated by simply averaging samples according to

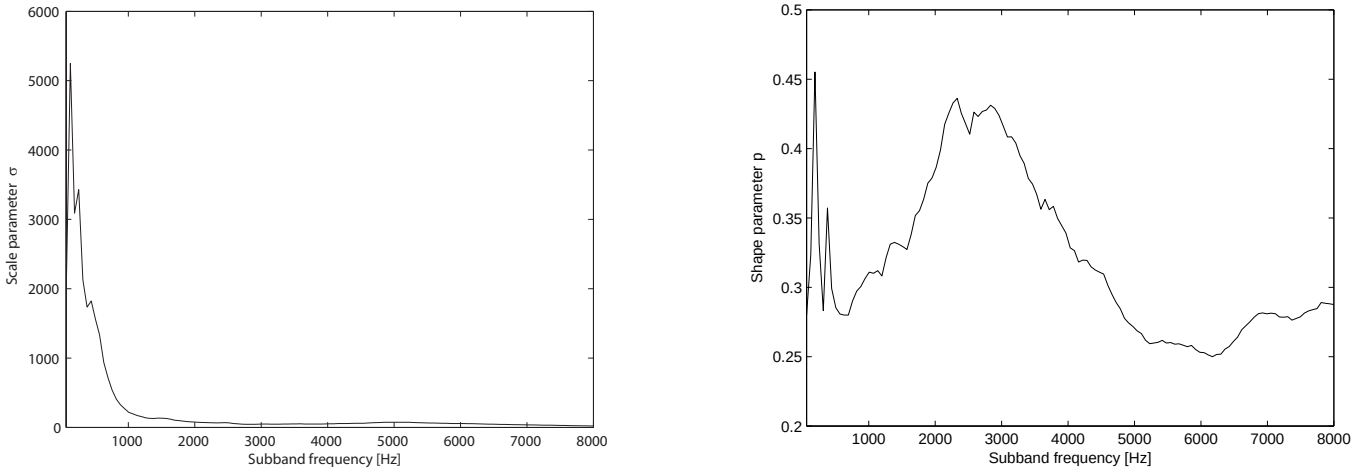
$$\text{kurt}(Y) = \frac{1}{N}\sum_{n=0}^{N-1}|Y_n|^4 - 3\left(\frac{1}{N}\sum_{n=0}^{N-1}|Y_n|^2\right)^2. \tag{12}$$

The kurtosis criterion does not require any explicit assumption as to the exact form of the pdf. Due to its simplicity, it is widely used as a measure of non-Gaussianity. The value calculated for kurtosis, however, can be strongly influenced by a few samples with a low observation probability. Hyvärinen and Oja [30] noted that negentropy is generally more robust in the presence of outliers than kurtosis. Hence, we adopt negentropy as our measure of choice, although we will also measure and report kurtosis values.

IV. Speech Modeling with the GG pdf

Subbands of speech can be precisely modeled by estimating the parameters of the GG pdf from training samples. From the trained parameters, insight can be gained into the statistical properties of human speech. Fig. 9 shows the scale parameter $\hat{\sigma}_{|Y|}$ and the shape parameter p calculated from the magnitude of subband components plotted as functions of frequency, where the number of the subbands is 256. The training samples used for estimating the GG pdf here were taken from clean speech data in the SSC2 development set [1].

It is clear from Fig. 9 that the scale parameter $\hat{\sigma}_{|Y|}$ becomes smaller at higher frequencies. The scale parameter $\hat{\sigma}_{|Y|}$ is related to the variance of $|Y|$, although not identical to it in the case that the ML method is used in its estimation. Fig. 9



(a) (b)

Fig. 9. The parameters of the GG pdf for frequency; (a) scale parameter $\hat{\sigma}_{|Y|}$ and (b) shape parameter p, where the sampling frequency is 16 kHz.

indicates that the magnitude at lower frequencies varies more than that at higher frequencies. Moreover, the GG pdfs trained with actual speech data are super-Gaussian with $p < 2$ in all subbands; they are in fact *super-Laplacian* with $p < 1$ in all subbands. As mentioned previously, the kurtosis is a measure of the super-Gaussianity of a pdf. It is therefore of interest to examine the behavior of kurtosis of the GG pdf. As demonstrated in Appendix A, the latter can be expressed as

$$\text{kurt}(Y_{gg}) = \hat{\sigma}^4 \left\{ \frac{\Gamma(1/p)\,\Gamma(5/p)}{\Gamma^2(3/p)} - 3 \right\}. \tag{13}$$

Fig. 10 shows a plot of kurtosis values as a function of frequency. In Fig. 10, a solid line indicates the kurtosis of the GG pdf calculated with (13) and a broken line presents the empirical kurtosis computed with (12). It is clear from Fig. 10 that the GG pdf can also model the kurtosis of speech, which would make the negentropy criterion more robust for outliers than the empirical kurtosis. It is also clear from Fig. 10 that kurtosis becomes smaller at higher frequencies, which indicates that the pdf of lower frequency components are more super-Gaussian than those of higher frequency components.

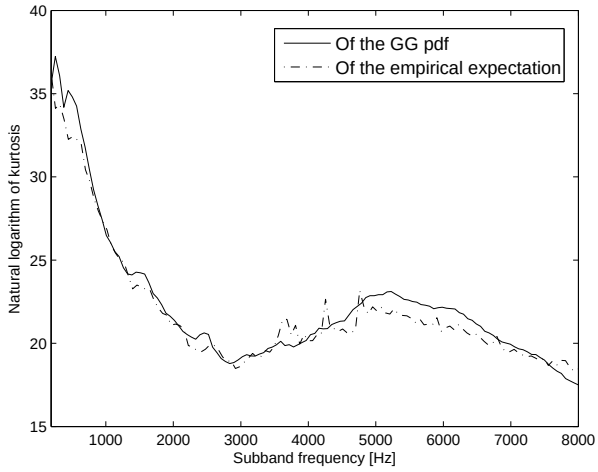


Fig. 10. Kurtosis for frequency, where the sampling frequency is 16 kHz.

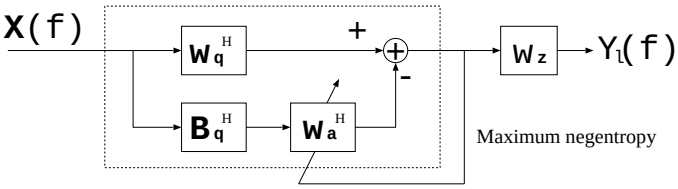


Fig. 11. Schematic of a generalized sidelobe canceling (GSC) beamformer for an active source.

V. Beamforming and Post-Filtering

Consider a subband beamformer in GSC configuration [4, §13.3.7] with a post-filter, as shown in Fig. 11. The output of a beamformer for a given subband can be expressed as

$$Y_t = \left(\mathbf{w}_{\mathrm{q}} - \mathbf{B}\mathbf{w}_{\mathrm{a}}\right)^H \mathbf{X}_t, \tag{14}$$

where $\mathbf{w}_{\mathrm{q}}$ is the *quiescent weight vector* for a source, $\mathbf{B}$ is the *blocking matrix*, $\mathbf{w}_{\mathrm{a}}$ is the *active weight vector*, and $\mathbf{X}_t$ is the input subband *snapshot vector* at frame t.

In keeping with the GSC formalism, $\mathbf{w}_{\mathrm{q}}$ is chosen to give unity gain in the desired *look direction* [4, §13.3.7]; i.e., to satisfy a *distortionless constraint*. The blocking matrix $\mathbf{B}$ is chosen to be orthogonal to $\mathbf{w}_{\mathrm{q}}$, such that $\mathbf{B}^H \mathbf{w}_{\mathrm{q}} = \mathbf{0}$. The blocking matrix can be calculated with an orthogonalization technique such as the modified Gram-Schmidt method, QR decomposition or singular value decomposition [39]. In this work, we used the modified Gram-Schmidt orthogonalization technique. The orthogonality implies that the distortionless constraint will be satisfied for any choice of $\mathbf{w}_{\mathrm{a}}$. While the active weight vector $\mathbf{w}_{\mathrm{a}}$ is typically chosen to minimize the variance of the beamformer's outputs, here we will develop an optimization procedure to find that $\mathbf{w}_{\mathrm{a}}$ which maximizes the negentropy $J(Y)$ described in Section III.

In order to calculate the negentropy, the variance of the

beamformer outputs Y is needed. Substituting (14) into the definition $\sigma_Y^2 = \mathcal{E}\{Y\,Y^*\}$ of variance, we find

$$\sigma_Y^2 = (\mathbf{w}_\mathrm{q} - \mathbf{B}\mathbf{w}_\mathrm{a})^H\,\mathbf{\Sigma_X}\,(\mathbf{w}_\mathrm{q} - \mathbf{B}\mathbf{w}_\mathrm{a})\,, \tag{15}$$

where $\mathbf{\Sigma_X} = \mathcal{E}\{\mathbf{X}\mathbf{X}^H\}$ is the covariance matrix of the input snapshot vectors.

Maximizing the negentropy criterion yields a weight vector $\mathbf{w}_\mathrm{a}$ capable of canceling interferences that leak through the sidelobes.

Zelinski post-filtering is performed on the output of the beamformer. The transfer function of the Zelinski post-filter can be expressed as

$$w_{z,t} = \frac{\frac{2}{M(M-1)}\left|\sum_{k=1}^{M-1}\sum_{l=k+1}^{M}\hat{\phi}_{kl,t}\right|}{\frac{1}{M}\sum_{k=1}^{M}\hat{\phi}_{kk,t}} \tag{16}$$

where $\hat{\phi}_{kk,t}$ is the auto-spectral density of the time-aligned input at microphone k and $\hat{\phi}_{kl,t}$ is the cross-spectral density (CSD) at microphone k and l. The estimation of a desired signal can be improved by averaging the CSDs [26]. The final output of the beamformer and post-filter combination is

$$Y_{l,t} = w_{z,t}\,Y_t = w_{z,t}\,(\mathbf{w}_\mathrm{q} - \mathbf{B}\mathbf{w}_\mathrm{a})^H\,\mathbf{X}_t. \tag{17}$$

For the experiments described in Section VII, subband analysis and synthesis were performed with a uniform DFT filter bank based on the modulation of a single prototype impulse response [40], which was designed to minimize each aliasing term individually. Beamforming in the subband domain has the considerable advantage that the active sensor weights can be optimized for each subband independently, which provides a tremendous computational saving with respect to a time-domain filter-and-sum beamformer with filters of the same length on the output of each sensor.

In conventional beamforming, *diagonal loading* is often applied in order to penalize large active weight vectors, and thereby improve robustness by inhibiting the formation of excessively large sidelobes [4, §13.3.8]. Such a regularization term can be applied in the present instance by defining the modified optimization criterion

$$\mathcal{J}(Y;\alpha) = J(Y) - \alpha\|\mathbf{w}_\mathrm{a}\|^2 \tag{18}$$

for some real $\alpha > 0$.

A. Estimation of Active Weights under the Γ pdf

Here we describe the formulae necessary for estimating the active weight vectors under the Γ pdf. Substituting (2) and (11) into (10), we can express the negentropy as

$$J(Y) = \log\left|\sigma_Y^2\right| + 2\,(1+\log 2\pi) + \frac{1}{T}\sum_{t=0}^{T-1}\log p_Y(Y_t), \tag{19}$$

where T is the number of frames used for weight vector adaptation. We maximize the objective function which is the sum of the negentropy and the negative regularization term. In the absence of a closed-form solution for the $\mathbf{w}_\mathrm{a}$ maximizing the negentropy (19), we resorted to the *conjugate gradients* method [41, §1.6].

By substituting (19) into (18) and taking the partial derivative on both sides, we obtain the gradient function,

$$\begin{aligned}\frac{\partial\mathcal{J}(Y;\alpha)}{\partial{\mathbf{w}_\mathrm{a}}^*} &= \frac{\partial J(Y;\alpha)}{\partial{\mathbf{w}_\mathrm{a}}^*} - \alpha\mathbf{w}_\mathrm{a} \\ &= \frac{1}{|\sigma_Y^2|}\frac{\partial|\sigma_Y^2|}{\partial{\mathbf{w}_\mathrm{a}}^*} + \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{p_Y(Y_t)}\frac{\partial p_Y(Y_t)}{\partial{\mathbf{w}_\mathrm{a}}^*} - \alpha\mathbf{w}_\mathrm{a}\end{aligned} \tag{20}$$

where

$$\frac{\partial|\sigma_Y^2|}{\partial{\mathbf{w}_\mathrm{a}}^*} = \frac{1}{T}\sum_{t=0}^{T-1}\left\{-\mathbf{B}^H\mathbf{X}_t Y_t^*\right\}. \tag{21}$$

Equations (20) and (21) are sufficient to implement a numerical optimization algorithm, whereby the negentropy $J(Y)$ can be maximized. The details of the numerical optimization algorithm are described in Appendix B.

B. Estimation of Active Weights under the Generalized Gaussian pdf

1) Parameter optimization 1: Unlike the pdfs that can be expressed as Meijer G-functions, the GG pdf cannot be readily extended from the univariate to the multi-variate. Hence, we use the magnitude of the beamformer's output as the r.v. for calculating the entropy. By substituting (5) and (11) into (10), we arrive at the following expression for negentropy

$$J(Y) = \log\left|\sigma_Y^2\right| + 2\,(1+\log 2\pi) - H_\mathrm{GG}(|Y|). \tag{22}$$

In order to apply the conjugate gradients algorithm, we must once more derive an expression for the gradient. By substituting (22) into (18) and taking the partial derivative on both sides while holding the shape parameter fixed, we obtain

$$\frac{\partial\mathcal{J}(Y;\alpha)}{\partial{\mathbf{w}_\mathrm{a}}^*} = \frac{1}{\sigma_Y^2}\frac{\partial\sigma_Y^2}{\partial{\mathbf{w}_\mathrm{a}}^*} - \frac{\partial H_\mathrm{GG}(|Y|)}{\partial{\mathbf{w}_\mathrm{a}}^*} - \alpha\mathbf{w}_\mathrm{a}, \tag{23}$$

where

$$\frac{\partial H_\mathrm{GG}(|Y|)}{\partial{\mathbf{w}_\mathrm{a}}^*} = \frac{1}{\hat{\sigma}_{|Y|}}\frac{\partial\hat{\sigma}_{|Y|}}{\partial{\mathbf{w}_\mathrm{a}}^*}. \tag{24}$$

Taking the derivative on both sides of (9), we find

$$\begin{aligned}\frac{\partial\hat{\sigma}_{|Y|}}{\partial{\mathbf{w}_\mathrm{a}}^*} = \frac{p}{T}\left[\frac{\Gamma(3/p)}{\Gamma(1/p)}\right]^{\frac{1}{2}} &\times \left[\frac{p}{T}\sum_{t=0}^{T-1}|Y_t|^p\right]^{\frac{1}{p}-1} \\ &\times \left[\sum_{t=0}^{T-1}|Y_t|^{p-1}\frac{\partial|Y_t|}{\partial{\mathbf{w}_\mathrm{a}}^*}\right],\end{aligned} \tag{25}$$

where the gradient of the magnitude at each frame is

$$\frac{\partial|Y_t|}{\partial{\mathbf{w}_\mathrm{a}}^*} = -\frac{1}{2|Y_t|}\mathbf{B}^H\mathbf{X}_t Y_t^*. \tag{26}$$

Based on (23) through (26), a numerical algorithm for optimizing the active weight vector can be implemented. The details of such an algorithm are given in Appendix B.

2) Parameter optimization 2: It is conceiveable that the entropy of the GG pdf for the complex valued r.v. could be approximated by assuming that the real and imaginary parts are independent. Under such an assumption, the differential entropy of the GG pdf can be expressed as

$$H(Y) \approx H_{\mathrm{r}}(Y_r) + H_{\mathrm{i}}(Y_i), \tag{27}$$

where Y_{r} is the real part of Y and Y_{i} is its imaginary part. Notice that the shape parameters for the real and imaginary parts must be trained individually.

Then, upon substituting (11) and (27) into (10) and adding the regularization term, we obtain the objective function

$$\begin{aligned}\mathcal{J}(Y;\alpha) = &\log\left|\sigma_Y^2\right| + 2\left(1+\log 2\pi\right) \\ &- H_{\mathrm{r}}(Y_{\mathrm{r}}) - H_{\mathrm{i}}(Y_{\mathrm{i}}) - \alpha\|\mathbf{w}_{\mathrm{a}}\|^2.\end{aligned} \tag{28}$$

In order to employ the gradient algorithm, we take the partial derivative of (28)

$$\begin{aligned}\frac{\partial\mathcal{J}(Y;\alpha)}{\partial {\mathbf{w}_{\mathrm{a}}}^*} &= \frac{1}{|\sigma_Y^2|}\frac{\partial|\sigma_Y^2|}{\partial {\mathbf{w}_{\mathrm{a}}}^*} - \frac{\partial H_{\mathrm{r}}(Y_{\mathrm{r}})}{\partial {\mathbf{w}_{\mathrm{a}}}^*} - \frac{\partial H_{\mathrm{i}}(Y_{\mathrm{i}})}{\partial {\mathbf{w}_{\mathrm{a}}}^*} - \alpha\mathbf{w}_{\mathrm{a}} \\ &= \frac{1}{|\sigma_Y^2|}\frac{\partial|\sigma_Y^2|}{\partial {\mathbf{w}_{\mathrm{a}}}^*} - \frac{1}{\hat{\sigma}_{|Y_{\mathrm{r}}|}}\frac{\partial\hat{\sigma}_{|Y_{\mathrm{r}}|}}{\partial {\mathbf{w}_{\mathrm{a}}}^*} - \frac{1}{\hat{\sigma}_{|Y_{\mathrm{i}}|}}\frac{\partial\hat{\sigma}_{|Y_{\mathrm{i}}|}}{\partial {\mathbf{w}_{\mathrm{a}}}^*} - \alpha\mathbf{w}_{\mathrm{a}}.\end{aligned} \tag{29}$$

We can readily calculate $\hat{\sigma}_{|Y_{\mathrm{r}}|}$ and $\hat{\sigma}_{|Y_{\mathrm{i}}|}$ in (29) based on (9). Each derivative can be obtained by replacing the magnitude $|Y_t|$ with an absolute value of the real part $|Y_{\mathrm{r},t}|$ or that of the imaginary part $|Y_{\mathrm{i},t}|$ in (25). The derivatives of the absolute values of the real and imaginary parts can be expressed, respectively, as

$$\frac{\partial|Y_{\mathrm{r},t}|}{\partial {\mathbf{w}_{\mathrm{a}}}^*} = -\frac{1}{2}\mathbf{B}^H\mathbf{X}_t \cdot \mathrm{sign}(Y_{\mathrm{r},t}) \tag{30}$$

and

$$\frac{\partial|Y_{\mathrm{i},t}|}{\partial {\mathbf{w}_{\mathrm{a}}}^*} = j\frac{1}{2}\mathbf{B}^H\mathbf{X}_t \cdot \mathrm{sign}(Y_{\mathrm{i},t}). \tag{31}$$

Equations (29) through (31) are used for the gradient algorithm.

VI. Simulation

Conventional beamforming algorithms determine the optimum weight vector that minimizes the variance of the beamformer's output,

$$\mathbf{w}^H\mathbf{\Sigma}_{\mathbf{X}}\mathbf{w}, \tag{32}$$

subject to the distortionless constraint in the look direction

$$\mathbf{w}^H\mathbf{d} = 1, \tag{33}$$

where $\mathbf{d}$ is the beam-steering vector. The well-known solution is called the minimum variance distortionless response (MVDR) beamformer [4, §13.3.1]. The weight vector of the MVDR beamformer can be expressed as

$$\mathbf{w}_{\mathrm{MVDR}} = \frac{\mathbf{\Sigma}_{\mathbf{X}}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{\Sigma}_{\mathbf{X}}^{-1}\mathbf{d}}. \tag{34}$$

Additional weight is typically added to the main diagonal of $\mathbf{\Sigma}_{\mathbf{X}}$ in order to avoid excessively large sidelobes in the beam pattern and the attendant nonrobustness [4, §13.3.7]. The MVDR beamfomers would attempt to null out any interfering signal, but are prone to the signal cancellation problem [5] whenever there is an interfering signal that is correlated with the desired signal. In realistic environments, interference signals are highly correlated with a target signal since the target signal is reflected from hard surfaces such as walls and tables. Therefore, the adaptation of the weight vector is usually halted whenever the desired source is active. Many techniques have been proposed in the literature to avoid signal cancellation. Perhaps the best-known of such algorithms is the robust beamformer in GSC configuration proposed by Hoshuyama *et al.* [11]. In the lower branch, their algorithm adaptively estimates a blocking matrix which cancels the signal correlated with the output from the upper branch. Accordingly, the reflections of a desired signal can be eliminated from the lower branch by the adaptive blocking matrix (ABM). The coefficient of the ABM has upper and lower limits in order to specify the maximum allowable target-direction error. Then, the active weight vectors are estimated so as to minimize the output of the beamformer. Since the ABM can remove the reflections from the lower branch, the signal cancellation problem is alleviated. However, the ABM cancels not only the reflections but also interference signals in the case that the output of the upper branch contains the interference components. In this case, their algorithm is unable to suppress the leaked interference signals. In reality, the interference signals are often present in the upper branch due to steering errors and *spatial aliasing* [4, §13.1.4]. Therefore, Hoshuyama's algorithm requires in some sense a trade-off between the avoidance of signal cancellation and suppression of the interference signals. This problem can be solved by simply halting the adaptation of the ABM and only updating the active weight vectors in the case of a high signal-to-noise ratio (SNR) [13]. Such a switching algorithm is based on SNR, however, and requires complicated rules which must generally be determined empirically.

Gannot et al. [8], [17], [18] proposed a transfer function GSC (TF-GSC) which incorporates transfer functions from the desired source to microphones into the upper branch of the GSC. The ratios of the transfer functions from the source to the microphone array are estimated with the least squares method when the desired signal is present. The quiescent vectors are calculated with the estimated ratios. The blocking matrices are then computed so as to satisfy the orthogonality condition with those quiescent weight vectors. Thus the leakage of the desired signal into the lower branch can be avoided. Their algorithm can estimate the ratios of the transfer function without source positions in acoustically stationary environments. It is difficult, however, to obtain stable solutions under non-stationary conditions. Although the algorithm proposed by Gannot et al. can be used in moderately reverberant environments, it does not reduce the amount of reverberation in the final signal [42].

E. Warsitz et al. proposed a generalized eigenvector (GEV) beamforming algorithm which constructs the blocking matrix based on the maximum SNR criterion [15]. They first calculate beamformer weights which satisfy the maximum SNR criterion. Secondly the orthogonal projection for constructing

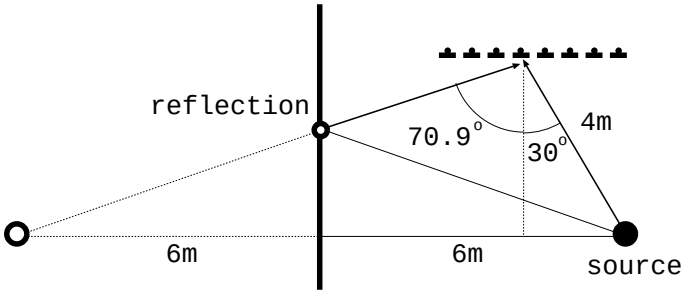


Fig. 12. Configuration of a source, sensors, and reflective surface for simulation.

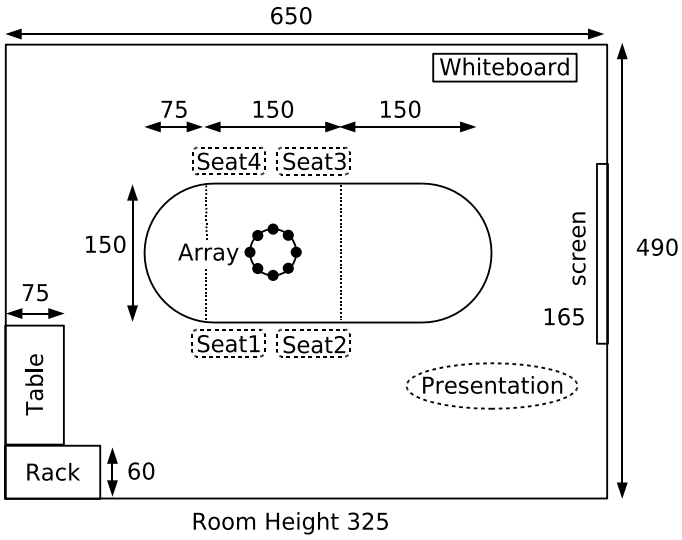


Fig. 15. The configuration of the meeting room (measurements in cm).

the blocking matrix is performed. Their algorithm estimates the transfer function from the source to the microphones indirectly. They demonstrated that their method could reduce signal distortion and noise more than the TF-GSC without post-filtering. It was also shown in [15] that their GEV beamforming algorithm can achieve almost the same noise suppression performance of the theoretical upper bound obtained by Hoshuyama's beamformer.

Based on the solutions mentioned above that have appeared in the literature, it could be argued that conventional robust beamforming algorithms have essentially addressed the problem of removing reflections that are highly correlated with the target signal in order to circumvent the signal cancellation problem.

In contrast to such conventional beamformers, the MN beamforming algorithm attempts not only to eliminate interference signals but also *strengthen* those reflections from the desired source, assuming the desired sound source is statistically independent of the other sources. 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 subband domain as long as it is shorter than the length of an analysis filter, and could thus be removed through a suitable choice of $\mathbf{w}_a$. Hence, the MN beamformer offers the possibility of steering both nulls and sidelobes; the former towards the undesired signal and its reflections, the latter towards reflections of the desired signal.

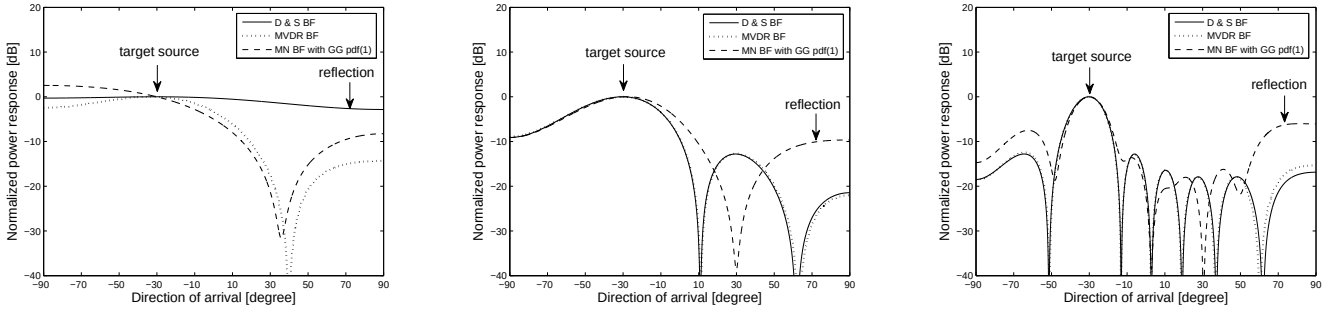
In order to verify that the MN beamforming algorithm forms sidelobes directed towards the reflection of a desired signal, we conducted experiments with a simulated acoustic environment. As shown in Fig. 12, we considered a simple configuration with a sound source, a reflective surface, and a linear array of eight microphones positioned with 10 cm inter-sensor spacing. Actual speech data were used as a source in this simulation, which was based on the *image method* [43]. White Gaussian noise was added to the output of each microphone to achieve a SNR of 0 dB. We assumed that the speed of sound is 343.74 meter per second and used a reflection coefficient of 0.7 for the wall. Fig. 13 shows beam patterns at f_s = 150 Hz, f_s = 650 Hz and f_s = 1600 Hz obtained with a delay-and-sum (D&S) beamformer, the MVDR beamformer and the MN beamforming algorithm with the GG pdf of the magnitude. The weights of the MVDR beamformer were optimized for isotropic (diffuse) noise in the simulation [44].

Given that a beam pattern shows the sensitivity of an array to plane waves, but the beam patterns in Fig. 13 were made with a near-field source and reflection, we also ran a second set of simulations in which the source and reflection were assumed to produce plane waves. The results of this second simulation are shown in Fig. 14. It is clear from these figures that the MN beamformer emphasizes reflections from the desired source. The MVDR beamformer optimized for the diffuse noise, on the other hand, tends to suppress such reflections. It is also apparent from Fig. 13 (a) and Fig. 14 (a) that MVDR and MN beamformers can suppress interference at low frequencies, while the suppression performance of the delay-and-sum beamformer is poor at low frequencies.

VII. Experiments

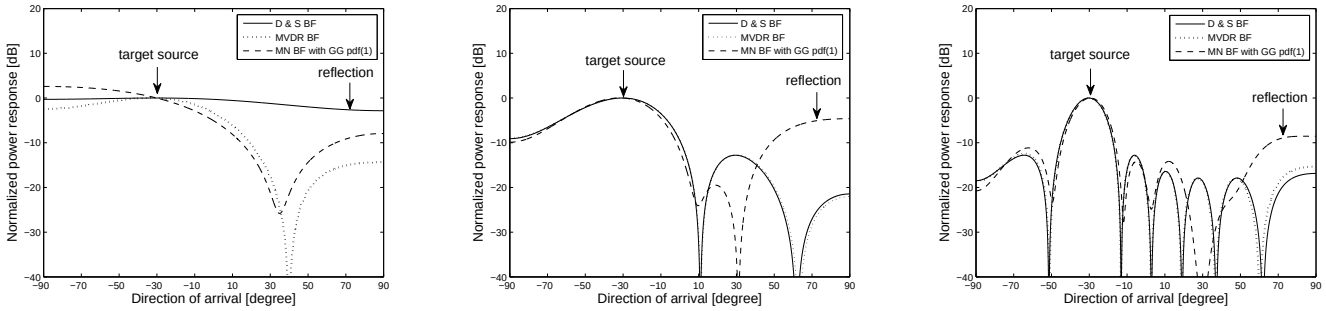
We performed far-field automatic speech recognition (ASR) experiments on the *Multi-Channel Wall Street Journal Audio Visual Corpus* (MC-WSJ-AV) collected by the *Augmented Multi-party Interaction* (AMI) project. The configuration of the meeting room is shown in Fig. 15; see Lincoln et al. [1] for the details of the data collection apparatus. The room size was 650 cm $\times$ 490 cm $\times$ 325 cm and the reverberation time T_{60} was approximately 380 milliseconds. In addition to being reverberant, the meeting room data collected includes background noise from computers and the building ventilation. Some recordings also contain audible noise from outside the meeting room, such as that generated by passing cars and speakers in an adjacent room.

The far-field speech data was recorded with a circular, eight-channel microphone array with a diameter of 20 cm. Additionally, a close-talking microphone was used for each speaker to capture the best possible signal as a reference. The sampling rate of the recordings was 16 kHz. As the data was recorded with real speakers in a realistic acoustic environment, the positions of the speakers' heads as well as the speaking volume varied even though the speakers were largely stationary. Indeed, it is exactly this behavior of real speakers that makes working with data from corpora such as MC-WSJ-



(a) (b) (c)

Fig. 13. Beam patterns produced by a delay-and-sum beamformer, the MVDR beamformer and the MN beamforming algorithm using a spherical wave assumption for (a) f_s = 150 Hz, (b) f_s = 650 Hz and (c) f_s = 1600 Hz.



(a) (b) (c)

Fig. 14. Beam patterns produced by a delay-and-sum beamformer, the MVDR beamformer and the MN beamforming algorithm using a plane wave assumption for (a) f_s = 150 Hz, (b) f_s = 650 Hz and (c) f_s = 1600 Hz.

AV so much more challenging than working with data that was played through a loudspeaker into a room, not to mention data that was *artificially convolved* with previously-measured impulse responses. In the *single speaker stationary* scenario of the MC-WSJ-AV, a speaker was asked to read sentences from six positions, four seated around the table in Seats 1-4 shown in Fig. 15, one standing at the white board, and one standing at the presentation screen.

The test set used for the experiments reported here contains recordings of 10 speakers where each speaker reads approximately 40 sentences taken from the 5,000 word vocabulary Wall Street Journal (WSJ) task. This provided a total of 352 utterances which correspond to 39.2 minutes of speech. There were a total of 11,598 word tokens in the reference transcriptions. The test set was disjoint from the training data set used to estimate the optimal scale and shape parameters.

As shown in [2] the directivity of the circular array at low frequencies is poor; this stems from the fact that for low frequencies, the wavelength is much longer than the aperture of the array. At high frequencies, the beam pattern is characterized by very large sidelobes; this is due to the fact that at high frequencies, the spacing between the elements of the array exceeds half of a wavelength, thereby causing spatial aliasing [4, §13.1.4].

Prior to beamforming, we first estimated the speaker's position with a speaker tracking system [45]. Based on the average speaker position estimated for each utterance, utterance-dependent active weight vectors $\mathbf{w}_a$ were estimated for a source. The active weight vectors for each subband were initialized to zero for estimation. Iterations of the conjugate gradients algorithm were run on the entire utterance until convergence was achieved.

Zelinski post-filtering [26] was performed after beamforming. The feature extraction of our ASR system was based on cepstral features estimated with a warped *minimum variance distortionless response* [46] (MVDR) spectral envelope of model order 30. Due to the properties of the warped MVDR, neither the Mel-filterbank nor any other filterbank was needed. The warped MVDR provides an increased resolution in low–frequency regions relative to the conventional Mel-filterbank. The MVDR also models spectral peaks more accurately than spectral valleys, which leads to improved robustness in the presence of noise. Front-end analysis involved extracting 20 cepstral coefficients per frame of speech and performing global cepstral mean subtraction (CMS) with variance normalization. The final features were obtained by concatenating 15 consecutive frames of cepstral features together, then performing a *linear discriminant analysis* (LDA) to obtain a feature of length 42. The LDA transformation was followed by a second global CMS, then a global semi-tied covariance (STC) transform [47].

The far-field ASR experiments reported here were conducted with a *word trace decoder* implemented along the lines suggested by Saon *et al.* [48]. The decoder is capable

of generating word lattices, which can then be optimized with weighted finite-state transducer (WFST) operations as in [49]; i.e., the raw lattice from the decoder is projected onto the output side to discard all arc information save for the word identities, and then compacted through epsilon removal, determinization, and minimization [50].

We used 30 hours of American WSJ and the 12 hours of Cambridge WSJ data in order to train a triphone acoustic model. The latter was necessary in order to provide coverage of the British accents for the speakers in the SSC development set [1]. Acoustic models estimated with two different HMM training schemes were used for the various decoding passes: conventional maximum likelihood (ML) HMM training [51, §12], and speaker-adapted training under a ML criterion (ML-SAT) [52]. Our baseline system was fully continuous with 1,743 codebooks and a total of 67,860 Gaussian components. The parameters of the GG pdf were trained with 43.9 minutes of speech data recorded with the CTM in the SSC development set. The training data set for the GG pdf contains recordings of 5 speakers.

We performed four decoding passes on the waveforms obtained with each of the beamforming algorithms described in prior sections. Each pass of decoding used a different acoustic model or speaker adaptation scheme. For all passes save the first unadapted pass, speaker adaptation parameters were estimated using the word lattices generated during the prior pass, as in [53]. A description of the four decoding passes follows:

1. Decode with the unadapted, conventional ML acoustic model.
2. Estimate vocal tract length normalization (VTLN) [54] parameters and constrained maximum likelihood linear regression parameters (CMLLR) [55] for each speaker, then redecode with the conventional ML acoustic model.
3. Estimate VTLN, CMLLR, and maximum likelihood linear regression (MLLR) [56] parameters for each speaker, then redecode with the conventional model.
4. Estimate VTLN, CMLLR, MLLR parameters for each speaker, then redecode with the ML-SAT model.

All passes used the full trigram LM for the 5,000 word WSJ task, which was made possible through the fast-on-the-fly composition algorithm described in [57].

Table I shows the word error rates (WERs) for every beamforming algorithm. As references, WERs in recognition experiments on speech data recorded with the single distant microphone (SDM) and CTM are also given. It is clear from Table I that every MN beamforming algorithm can provide better recognition performance than the simple delay-and-sum beamformer (D&S BF) which can be improved by Zelinski post-filtering (D&S BF with PF). It is also clear from Table I that MN beamforming with the GG pdf assumption which uses the magnitude in calculating the negentropy (MN BF with GG pdf (1)) achieves the best recognition performance. This is due to the fact that the GG pdf models the magnitudes of the subband samples of speech better than the other pdfs in that the shape parameter for each subband is estimated individually from training data. The recognition performance, however, did not improve for MN beamforming with the GG pdf when the real and imaginary parts of the subband components were assumed to be independent (MN BF with GG pdf (2)). We found it better to treat the subband components as spherically-invariant random processes (SIRPs) as in [2], [33] and are led to conclude that the real and imaginary parts are dependent as mentioned in [25]. Table I suggests that the Γ pdf assumption (MN BF with Γ pdf) can lead to better noise suppression performance to some extent. The reduction over the D&S BF with PF case, however, is limited because the Γ pdf cannot model the subband components of speech as precisely as the GG pdf which takes the magnitude as the r.v. We also performed recognition experiments on speech enhanced by the MVDR beamformer with Zelinski post-filtering, which is equivalent to the minimum mean-squared error beamformer (MMSE BF) [4, §13.3.5]. Table I demonstrates that the MVDR beamformer with post-filtering (MMSE BF) provides better recognition performance than D&S BF with PF. The MMSE beamformer would suppress the reflections of the desired signal. On the other hand, as demonstrated in Section VI, the MN beamforming algorithm can strengthen the target signal by using the reflections solely based on the maximum negentropy criterion. Note that the MVDR beamforming algorithms require speech activity detection in order to avoid signal cancellation. For the adaptation of the MVDR beamformer, we used the first 0.1 and last 0.1 seconds in each utterance, which contain only background noise. Table I also shows the recognition results obtained with the generalized eigenvector beamformer (GEV BF) proposed by E. Warsitz et al. [15]. It achieved slightly better recognition performance than the MMSE beamformer. In this task, the transfer function from the sound source to the microphone array changes in time due to movements of the speaker's head. Moreover, it is difficult to determine whether or not the signal observed at any given time contains both speech and noise components in each frequency bin, which are required to estimate the transfer function. Due to these difficulties, the performance of the GEV beamformer is limited in realistic environments. Once more, in contrast to conventional beamforming methods, our algorithm does not need to detect the start and end points of target speech since the proposed method can suppress noise and reverberation without the signal cancellation problem. It is worth noting that the best result of 13.2% in Table I is significantly less than half the word error rate reported elsewhere in the literature on this far-field ASR task [1].

We also examined the effect of the regularization term in equation (18). Table II shows WER as a function of the regularization parameter α, where we used the MN beamforming algorithm with the GG pdf of the magnitude r.v. We can see from the table that the regularization parameter $\alpha = 10^{-2}$ provided the lowest word error rate, although the impact of different values of α on recognition performance was slight. The regularization parameter α could be interpreted as an indicator of the sufficiency of the input data in estimating the active weight vector. Thus, the requirement of a small α may imply that the input data are not sufficiently reliable to completely determine the active weight vector due to, for example, steering errors.

We implemented each beamforming algorithm in C/C++

TABLE I

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

Beamforming Algorithm	Pass (%WER)			
	1	2	3	4
D&S BF	80.1	39.9	21.5	17.8
D&S BF with PF	79.0	38.1	20.2	16.5
MMSE BF	78.6	35.4	18.8	14.8
GEV BF	78.7	35.5	18.6	14.5
MN BF with Gamma pdf	75.6	34.9	19.8	15.8
MN BF with GG pdf (1)	75.1	32.7	16.5	13.2
MN BF with GG pdf (2)	79.0	37.2	20.0	16.7
SDM	87.0	57.1	32.8	28.0
CTM	52.9	21.5	9.8	6.7

TABLE II

WORD ERROR RATES AGAINST THE REGULARIZATION PARAMETER α.

α	Pass (%WER)			
	1	2	3	4
$\alpha = 0.0$	72.7	31.9	16.4	13.7
$\alpha = 10^{-3}$	73.9	32.2	16.6	13.6
$\alpha = 10^{-2}$	75.1	32.7	16.5	13.2
$\alpha = 10^{-1}$	76.2	32.5	17.5	13.5

and python. The computational cost of the MN beamforming algorithm (MN BF with GG pdf (1)) is approximately 2.6 times as much as that of the MMSE beamformer per frame on a machine with an Intel Core 2 DUO E6750/2.66GHz processor and 3.36 GB RAM.

VIII. CONCLUSIONS AND FUTURE WORK

In this work, we have proposed a novel beamforming algorithm based on maximizing negentropy. Our first investigations into the MN beamforming algorithm were based on acoustic simulations. These simulations were sufficient to demonstrate the MN beamforming algorithm could strengthen the desired signal by constructively adding reflections of the same. Moreover, the proposed method does not exhibit the signal cancellation problems typically seen in conventional beamformers. We also evaluated the Γ and GG pdfs in calculating the negentropy through a set of far-field automatic speech recognition experiments with data captured in realistic acoustic environments and spoken by real speakers. In these experiments, the MN beamforming algorithm with the GG pdf assumption proved to provide the best ASR performance.

We plan to develop an on–line version of the beamforming algorithm presented here. This on–line algorithm will be capable of adjusting the active weight vectors $\mathbf{w}_{\mathrm{a},i}$ with each new snapshot in order to track changes of speaker position and movements of the speaker's head during an utterance.

ACKNOWLEDGEMENT

We would like to thank Prof. Hervé Bourlard for giving us the opportunity to study about far-field speech recognition.

APPENDIX

A. The r-th moment and kurtosis of the GG pdf

In this section, we derive two useful statistics of the GG pdf, the r-th moment and kurtosis.

The rth moment of the GG pdf can be expressed as

$$\mathcal{E}\{y^r\} = \frac{1}{2\Gamma(1+1/p)A(p,\hat{\sigma})}\int_{-\infty}^{\infty} y^r \exp\left[-\frac{|y|^p}{A(p,\hat{\sigma})}\right] dy. \quad (35)$$

Since the GG pdf is an even function about the mean, we can rewrite (35) as

$$\mathcal{E}\{y^r\} = \frac{1}{\Gamma(1+1/p)A(p,\hat{\sigma})}\int_{0}^{\infty} y^r \exp\left[-\frac{y^p}{A^p(p,\hat{\sigma})}\right] dy. \quad (36)$$

Upon defining

$$v = \frac{y^p}{A^p(p,\hat{\sigma})},$$

from which it follows

$$\frac{dv}{dy} = \frac{py^{p-1}}{A^p(p,\hat{\sigma})},$$

then (36) can be solved as

$$\begin{aligned}\mathcal{E}\{y^r\} &= \frac{A^r(p,\hat{\sigma})}{p\,\Gamma(1+1/p)}\int_0^{\infty} v^{\frac{r+1}{p}-1}\, e^{-v}\, dv \\ &= \frac{A^r(p,\hat{\sigma})}{p\,\Gamma(1+1/p)}\,\Gamma\left(\frac{r+1}{p}\right). \quad (37)\end{aligned}$$

By substituting the second and fourth moments obtained from Equation (37), the kurtosis of the GG pdf can now be expressed as

$$\mathrm{kurt}\,(Y_{gg}) = \frac{A(p,\hat{\sigma})^4}{p\Gamma(1+1/p)}\Gamma\,(5/p) - 3\left\{\frac{A(p,\hat{\sigma})^2}{p\Gamma(1+1/p)}\Gamma\,(3/p)\right\}^2. \quad (38)$$

As $p\Gamma(1+1/p) = \Gamma(1/p)$, Eqn. (38) can be simplified to

$$\mathrm{kurt}\,(Y_{gg}) = \hat{\sigma}^4\left\{\frac{\Gamma(1/p)\,\Gamma(5/p)}{\Gamma^2(3/p)} - 3\right\}. \quad (39)$$

B. The implementation of the optimization algorithm

Here we describe a nonlinear conjugate gradient method for our beamforming algorithm. Our goal is to find the active weight vector which provides the maximum negentropy. However, gradient algorithms are generally used to find the local minimum of a function [38, §1.6]. Accordingly, we explain how to find the local minimum of the negative of (18) with a conjugate gradient algorithm, which is equivalent to seeking the local maximum of (18).

The conjugate algorithms proceed as a succession of line minimizations. The sequence of *conjugate directions* is used to approximate the curvature of a cost function in the neighborhood of the minimum.

Expressing the objective function as $\mathcal{I}(\mathbf{w}_{\mathrm{a}}{}^*) = -\mathcal{J}(Y;\alpha)$, we can calculate the initial search direction as that opposite to the gradient according to

$$\Delta\mathbf{w}_{\mathrm{a}(0)}^* = -\frac{\partial\mathcal{I}(\mathbf{w}_{\mathrm{a}(0)}^*)}{\partial\mathbf{w}_{\mathrm{a}}{}^*},$$

where the required partial derivative is specified by one of (20), (23) or (29). A line search is performed in that direction and a step size is optimized as follows:

$$\begin{aligned}\beta_{(0)} &:= \mathrm{argmin}_\beta\, \mathcal{I}(\mathbf{w}_{\mathrm{a}}{}^* + \beta\Delta\mathbf{w}_{\mathrm{a}(0)}^*) \text{ and} \\ \mathbf{w}_{\mathrm{a}(1)}^* &= \mathbf{w}_{\mathrm{a}(0)}^* + \beta_{(0)}\Delta\mathbf{w}_{\mathrm{a}(0)}^*,\end{aligned}$$

where the initial active weight vector is set to zero in this work.

After the first iteration, the following steps constitute one iteration of searching the minimum along a subsequent conjugate direction $\Lambda \mathbf{w}_{\mathrm{a}(n)}^*$, where $\Lambda \mathbf{w}_{\mathrm{a}(0)}^* = \Delta \mathbf{w}_{\mathrm{a}(0)}^*$:

1. Calculate the gradient of the objective function
$$\Delta \mathbf{w}_{\mathrm{a}(n)}^* = -\frac{\partial \mathcal{I}(\mathbf{w}_{\mathrm{a}(n)}^*)}{\partial \mathbf{w}_{\mathrm{a}}{}^*}.$$
2. Compute the modified Polak-Ribière formula
$$\gamma_{(n)} = Re\left\{ \frac{\Delta \mathbf{w}_{\mathrm{a}(n)}^T \left(\Delta \mathbf{w}_{\mathrm{a}(n)}^* - \Delta \mathbf{w}_{\mathrm{a}(n-1)}^* \right)}{\Delta \mathbf{w}_{\mathrm{a}(n-1)}^T \Delta \mathbf{w}_{\mathrm{a}(n-1)}^*} \right\},$$
where $(\cdot)^T$ denotes the transpose operation.
3. Update the conjugate direction
$$\Lambda \mathbf{w}_{\mathrm{a}(n)}^* = \Delta \mathbf{w}_{\mathrm{a}(n)}^* + \gamma_{(n)} \Lambda \mathbf{w}_{\mathrm{a}(n-1)}^*.$$
4. Perform the line search and optimize the step size
$$\beta_{(n)} = \mathrm{argmin}_\beta \, \mathcal{I}(\mathbf{w}_{\mathrm{a}(n)}^* + \beta \Lambda \mathbf{w}_{\mathrm{a}(n)}^*). \tag{40}$$
5. Update the estimate of the active weight vector
$$\mathbf{w}_{\mathrm{a}(n+1)}^* = \mathbf{w}_{\mathrm{a}(n)}^* + \beta_{(n)} \Lambda \mathbf{w}_{\mathrm{a}(n)}^*.$$

In each step, the line search is repeated until

$$\mathrm{Re}\left\{ \Delta \mathbf{w}_{\mathrm{a}(n)} \cdot \Lambda \mathbf{w}_{\mathrm{a}(n)}^* \right\} < \mathrm{tol} \, |\Delta \mathbf{w}_{\mathrm{a}(n)}| \, |\Lambda \mathbf{w}_{\mathrm{a}(n)}|. \tag{41}$$

where tol indicates the accuracy of the line search. We set $\mathrm{tol} = 0.001$ in our experiments. The convergence properties of the numerical search were not significantly altered by changing the method used to calculate $\gamma_{(n)}$, nor by adjusting the accuracy of the line search. Applying a more accurate model for the pdf of the subband samples of speech had a larger effect on the speed of convergence than any adjustment of the parameters of the conjugate gradients search.

REFERENCES

[1] 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.

[2] K. Kumatani, T. Gehrig, U. Mayer, E. Stoimenov, J. McDonough, and M. Wölfel, "Adaptive beamforming with a minimum mutual information criterion," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 2527–2541, 2007.

[3] H. K. Maganti, D. Gatica-Perez, and I. McCowan, "Speech enhancement and recognition in meetings with an audio-visual sensor array," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 2257–2269, 2007.

[4] M. Wölfel and J. McDonough, *Distant Speech Recognition*. New York: Wiley, 2009.

[5] B. Widrow, K. M. Duvall, R. P. Gooch, and W. C. Newman, "Signal cancellation phenomena in adaptive antennas: Causes and cures," *IEEE Transactions on Antennas and Propagation*, vol. AP-30, pp. 469–478, 1982.

[6] S. Nordholm, I. Claesson, and B. Bengtsson, "Adaptive array noise suppression of handsfree speaker input in cars," *IEEE Transactions on Vehicular Technology*, vol. 42, pp. 514–518, 1993.

[7] W. Herbordt and W. Kellermann, "Adaptive beamforming for audio signal acquisition," in *Adaptive Signal Processing: Applications to Real-World Problems*, J. Benesty and Y. Huang, Eds. Berlin, Germany: Springer, 2003, pp. 155–194.

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

[9] I. Claesson and S. Nordholm, "A spatial filtering approach to robust adaptive beaming," *IEEE Transactions on Antennas and Propagation*, vol. 19, pp. 1093–1096, 1992.

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

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

[12] N. Grbić, "Optimal and adaptive subband beamforming," Ph.D. dissertation, Blekinge Institute of Technology, 2001.

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

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

[15] E. Warsitz, A. Krueger, and R. Haeb-Umbach, "Speech enhancement with a new generalized eigenvector blocking matrix for application in a generalized sidelobe canceller," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, U.S.A, 2008.

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

[17] S. Gannot, David, Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, pp. 1614–1626, 2001.

[18] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Transactions Speech and Audio Processing*, vol. 12, pp. 561–571, 2004.

[19] S. Y. Low, S. Nordholm, and R. Togneri, "Convolutive blind signal separation with post-processing," *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 539–548, 2004.

[20] H. Buchner, R. Aichner, and W. Kellermann, "Blind source seperation for convolutive mixtures: A unified treatment," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds. Boston: Kluwer Academic, 2004, pp. 255–289.

[21] H. Saruwatari, T. Kawamura, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ica and beamforming," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 666–678, 2006.

[22] P. Smaragdis, "Efficient blind separation of convolved sound mixtures," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, U.S.A, 1997.

[23] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Transactions Speech Audio Processing*, vol. 10, no. 6, pp. 352–362, September 2002.

[24] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Transactions Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, Sept. 2005.

[25] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 1741–1752, 2007.

[26] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with post-filtering," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 240–259, 1998.

[27] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 709–716, 2003.

[28] S. Leukimmiatis, D. Dimitriadis, and P. Maragos, "An optimum microphone array post-filter for speech application," in *Proc. Interspeech-ICSLP*, Pittsburgh, PA, U.S.A, 2006.

[29] H. Yoon and H. Ko, "Microphone array post-filter using inputoutput ratio of beamformer noise power spectrum," *Electronics Letters*, vol. 43, pp. 1003–1005, 2007.

[30] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, 2000.

[31] R. G. Gallager, *Information Theory and Reliable Communication*. New York: John Wiley & Sons, 1968.

[32] F. D. Neeser and J. L. Massey, "Proper complex random processes with applications to information theory," *IEEE Transactions Info. Theory*, vol. 39, no. 4, pp. 1293–1302, July 1993.

[33] H. Brehm and W. Stammler, "Description and generation of spherically invariant speech-model signals," *Signal Processing*, vol. 12, pp. 119–141, 1987.

[34] S. Wolfram, *The Mathematica Book*, 3rd ed. Cambridge: Cambridge University Press, 1996.

[35] K. Kokkinakis and A. K. Nandi, "Exponent parameter estimation for generalized gaussian probability density functions with application to speech modeling," *Signal Processing*, vol. 85, pp. 1852–1858, 2005.

[36] M. K. Varanasi and B. Aazhang, "Parametric generalized gaussian density estimation," *J. Acoust. Soc. Am.*, vol. 86, pp. 1404–1415, 1989.

[37] M. K. Varanasi, "Parameter estimation for the generalized gaussian noise model," Ph.D. dissertation, Rice University, 1987.

[38] D. P. Bertsekas, *Nonlinear Programming*. Belmont, Massachusetts: Athena Scientific, 1995.

[39] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore: The Johns Hopkins University Press, 1996.

[40] K. Kumatani, J. McDonough, S. Schacht, D. Klakow, P. N. Garner, and W. 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.

[41] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA, USA: Athena Scientific, 1995.

[42] E. A. P. Habets, "Single- and multi microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Eindhoven University of Technology, 2007.

[43] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acou stics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, April 1979.

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

[45] T. Gehrig, U. Klee, J. McDonough, S. Ikbal, M. Wölfel, and C. Fügen, "Tracking and beamforming for multiple simultaneous speakers with probabilistic data association filters," in *Proc. Interspeech*, 2006, pp. 2594–2597.

[46] M. Wölfel and J. McDonough, "Minimum variance distortionless response spectral estimation: Review and refinements," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 117–126, Sept. 2005.

[47] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions Speech and Audio Processing*, vol. 7, pp. 272–281, 1999.

[48] G. Saon, D. Povey, and G. Zweig, "Anatomy of an extremely fast LVCSR decoder," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 549–552.

[49] A. Ljolje, F. Pereira, and M. Riley, "Efficient general lattice generation and rescoring," in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 1251–1254.

[50] M. Mohri and M. Riley, "Network optimizations for large vocabulary speech recognition," *Speech Communication*, vol. 28, no. 1, pp. 1–12, 1999.

[51] J. Deller, J. Hansen, and J. Proakis, *Discrete-Time Processing of Speech Signals*. New York: Macmillan Publishing, 1993.

[52] T. Anastasakos, J. McDonough, R. Schwarz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP*, 1996, pp. 1137–1140.

[53] L. Uebel and P. Woodland, "Improvements in linear transform based speaker adaptation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001.

[54] L. Welling, H. Ney, and S. Kanthak, "Speaker adaptive modeling by vocal tract normalization," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 415–426, 2002.

[55] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, 1998.

[56] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, April 1995.

[57] J. McDonough, E. Stoimenov, and D. Klakow, "An algorithm for fast composition of weighted finite-state transducers," in *Proc. IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2007.