IDIAP RESEARCH REPORT

Maximum Negentropy Beamforming

Kenichi Kumatani [a] John McDonough [b]
Dietrich Klakow [b] Philip N. Garner [a]
Weifeng Li [a]

IDIAP–RR 08-07

April 2008

[a] IDIAP Research Institute, Martigny, Switzerland
[b] Spoken Language Systems at Saarland University in Saarbrücken, Germany

IDIAP Research Institute **www.idiap.ch**
Av. des Prés–Beudin 20 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 08-07

Maximum Negentropy Beamforming

Kenichi Kumatani John McDonough Dietrich Klakow Philip N. Garner
Weifeng Li

April 2008

Abstract.

In this paper, we address an adaptive beamforming application based on the capture of far-field speech data from a single speaker in a real meeting room. After the position of a 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 adaptive beamforming algorithms. We demonstrate this fact through experiments on acoustic simulations. Moreover, 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), 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%.

1 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. Adaptive 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, §6.7.3]. Subject to this *distortionless 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, the adaptation of the active weight vector is typically halted whenever the desired source is active.

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. In other words, it represents the degree of super-Gaussianity of a distribution [6]. The pdf of speech is in fact super-Gaussian [2, 7, 8], 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 adaptive beamforming algorithms. For calculating negentropy, we consider the Γ and the 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 [9].

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 by the European Union integrated project *Augmented Multi-party Interaction* (AMI) [1].

The balance of this work is organized as follows. We describe the super-Gaussian pdfs which are used for calculating the negentropy in Section 2. In particular, Section 2 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 3 reviews the definition of entropy and negentropy. In Section 4, we discuss our maximum negentropy beamforming criterion and then derive the objective functions for estimating the active weight vectors. Section 5 illustrates the speech distribution modeled with the GG pdf. In Section 6, we demonstrate that the proposed beamforming algorithm has no signal cancellation problem through acoustic simulations. In Section 7, we describe the results of far-field automatic speech recognition experiments. Finally, in Section 8, we present our conclusions and plans for future work.

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

Here we present theoretical arguments and empirical evidence that subband samples of speech, like nearly all other information bearing signals, are *not* Gaussian-distributed [6]. Hence, we are led to consider the use of super-Gaussian pdfs to model the subband samples of speech, as well as to calculate the negentropy of outputs of a GSC.

The entire field of *independent component analysis* (ICA) is founded on the assumption that all signals of real interest are *not* Gaussian-distributed. A concise and very readable argument for the validity of this assumption is given by Hyvärinen and Oja [6]. Briefly, their 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

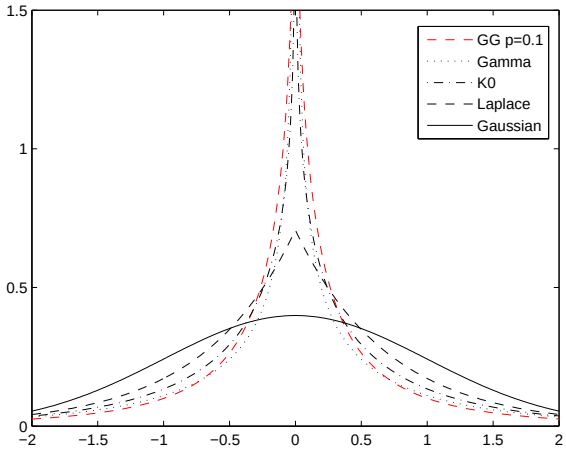


Figure 1: Plot of the likelihood of the super-Gaussian and Gaussian pdfs.

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. As discussed in Section 3, entropy is the basic measure of information in *information theory* [10]. It is well known that a Gaussian r.v. has the highest entropy of all r.v.s with a given variance [10, Thm. 7.4.1], which also holds for complex Gaussian r.v.s [11, Thm. 2]. Hence, a Gaussian r.v. is, in some sense, the least *predictable* of all r.v.s., which is why the Gaussian pdf is most often associated with *noise*. Interesting signals contain structure that makes them more predictable than Gaussian r.v.s. Hence, if an interesting 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, 7, 8]. Noise, on the other hand, is typically Gaussian-distributed. In fact, the pdf of the sum of super-Gaussian variables gets closer to Gaussian. Thus, a mixture signal which consists of many interference signals can be expected to be Gaussian-distributed. Based on these facts, we might remove interference signals and extract a target signal by making the pdf of the beamformer's output as super-Gaussian as possible.

A plot of the likelihood of the Gaussian and four super-Gaussian univariate pdfs considered is provided in Fig. 1, where the likelihood of the generalized Gaussian (GG) pdf is calculated with the shape parameter $p = 0.1$. From the figure, it is clear that the Laplace, K_0, Γ, and GG with $p = 0.1$ 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 the histogram of the real parts of subband components at $f_s = 800$ Hz. To generate these histograms, we used 43.9 minutes of clean speech recorded with the CTM in the development set of the Speech Separation Challenge, Part 2 (SSC2) [1]. Fig. 2 also presents the likelihoods of the pdfs. In Fig. 2, the parameters of the GG pdf are 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 speech.

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

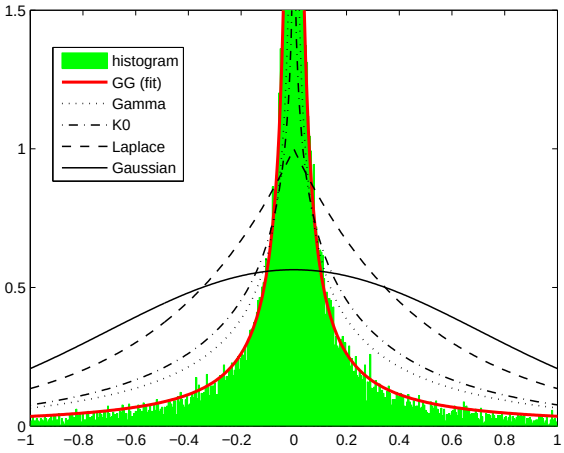


Figure 2: Histogram of real parts of subband components and the likelihood of pdfs.

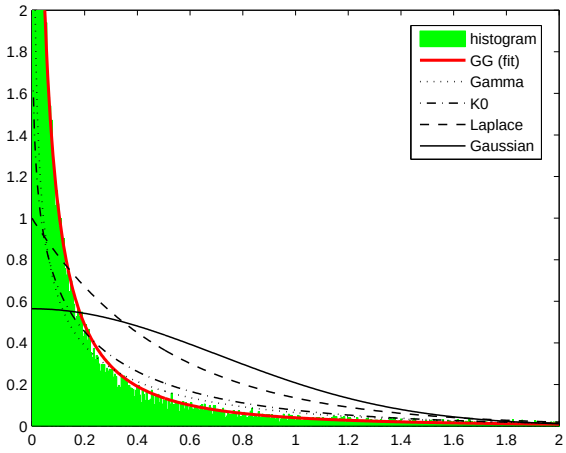


Figure 3: Histogram of magnitude in the subband domain and the likelihood of pdfs.

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 reverberated speech in the subband domain. In order to produce reverberated 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 reverberated speech in Fig. 7, respectively. We can again see from Fig. 6 and Fig. 7 that the pdfs of corrupted speech have the less probability mass around the mean and less probability mass in the tail, but once 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.

2.1 Super-Gaussian pdf derived from the Meijer G-function

Brehm and Stammler stated in [12] that it was useful to assume that the Laplace, K_0, and Γ pdfs belonged to the class of the *spherically-invariant random processes* (SIRPs) for two principal reasons. Firstly, this implies that multivariate pdfs of all orders can be readily derived from the univariate pdf using the theory of *Meijer G-function* based solely on the knowledge of the covariance matrix of the random vectors. Secondly, such variates can be extended to the case of complex r.v.s, which is essential for our current development.

We used the Γ pdf here since 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.

The differential entropy for the Γ pdf cannot be expressed in closed form, either. We must,

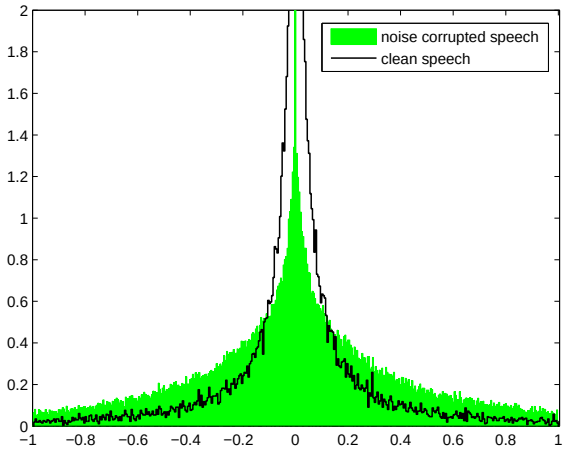


Figure 4: Histograms of clean speech and noise corrupted speech in the subband domain.

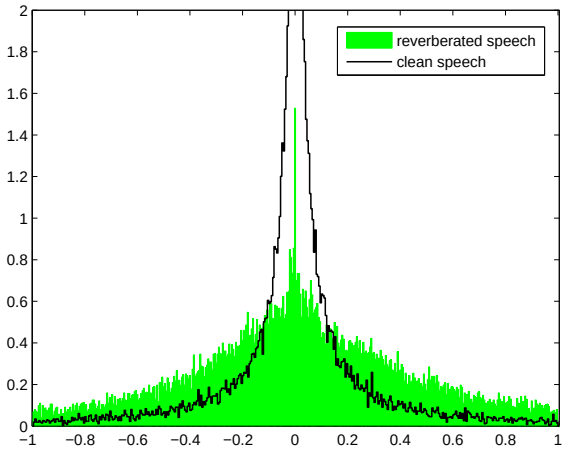


Figure 5: Histograms of clean speech and reverberated speech in the subband domain.

therefore, replace the exact differential entropy with the *empirical differential entropy*

$$H(Y) = -\mathcal{E}\left\{\log p_Y(Y)\right\} \approx -\frac{1}{T}\sum_{t=0}^{T-1}\log p_Y(Y_t). \tag{1}$$

2.2 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 parameter space [12]. 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 blind source separation (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

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

where

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

In (3), $\Gamma(.)$ is the gamma function and p is the shape parameter, which controls how fast the tail of the pdf decays. 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[1].

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

[1]Equation (2) is defined over the interval (-∞, +∞). Precisely speaking, the double-sided pdf (2) should be modified in order to model magnitude whose value is always positive. It is easily done by multiplying both sides of (2) by a factor of two and redefining the interval as [0, +∞). However, such modifications are not necessary in our algorithm because the double factor for the normalization is constant in the log-likelihood domain and has no effect on the gradient algorithm.

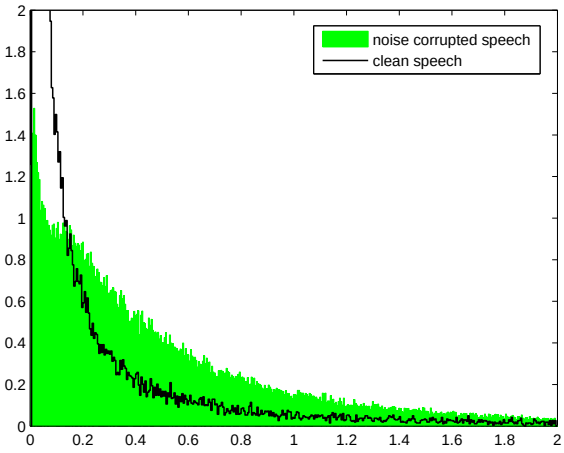


Figure 6: Histograms of magnitude of clean speech and noise corrupted speech in the subband domain.

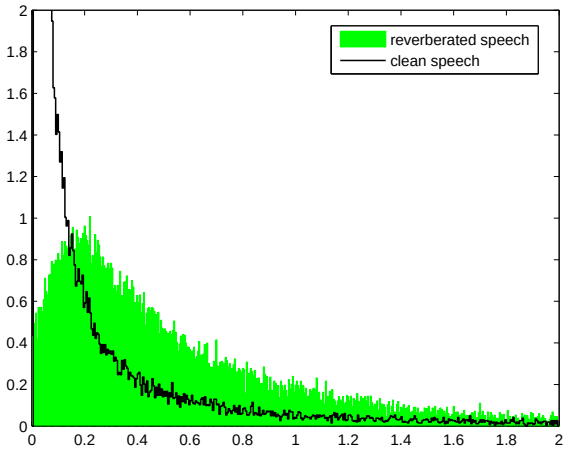


Figure 7: Histograms of magnitude of clean speech and reverberated speech in the subband domain.

The differential entropy of the GG pdf for the real-valued r.v. y is obtained with the help of *Mathematica* [13] as

$$\begin{aligned} H_{\mathrm{GG}}(y) &= -\int_{-\infty}^{+\infty} p_{\mathrm{gg}}(\xi)\log p_{\mathrm{gg}}(\xi)d\xi \\ &= \frac{1}{p}+\log\left[2\Gamma(1+1/p)A(p,\hat{\sigma})\right] \end{aligned} \tag{4}$$

The shape parameter p is trained with the method described in Section 2.3.

2.3 Methods for Estimating Scale and Shape Parameters

Among several methods for estimating the shape parameter p of the GG pdf [14][15], the moment and maximum likelihood (ML) methods are arguably the most straightforward. In this work, we use the moment method in order to initialize the parameters of the GG pdf and then update them with the ML estimate [15]. The shape parameters are estimated from training samples offline and are held fixed during the adaptation of the active weight vector. The shape parameters for each subband are estimated independently, as the optimal pdf is frequency-dependent.

For a set $\{Y_t\}$ of training data consisting of complex subband samples of speech, the log-likelihood function under the GG pdf can be expressed as

$$\begin{aligned} l(\hat{\sigma},p;y) = &-N\log\left\{2\Gamma(1+1/p)A(p,\hat{\sigma})\right\} \\ &-\frac{1}{A(p,\hat{\sigma})^p}\sum_{n=0}^{N-1}|y_n|^p, \end{aligned} \tag{5}$$

where N is the number of training samples. The parameters $\hat{\sigma}$ and p can be obtained by solving the following equations:

$$\frac{\partial l(\hat{\sigma},p;y)}{\partial\hat{\sigma}} = -\frac{N}{\hat{\sigma}}+\frac{p}{\hat{\sigma}^{p+1}}\left[\frac{\Gamma(1/p)}{\Gamma(3/p)}\right]^{-\frac{p}{2}}\sum_{n=0}^{N-1}|y_n|^p = 0, \tag{6}$$

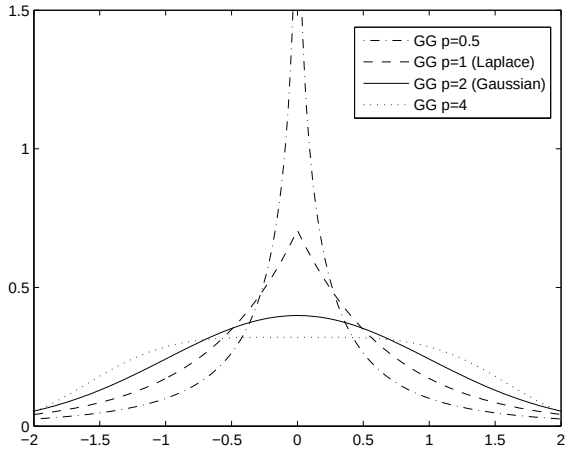


Figure 8: Plot of the likelihood of the generalized Gaussian (GG) pdfs.

$$\begin{aligned}\frac{\partial l(\hat{\sigma},p;y)}{\partial p} =& Na(p) - \sum_{n=0}^{N-1}\left(\frac{|y_n|}{A(p)}\right)^p \\ &\times\left[\log\left\{\frac{|y_n|}{A(p)}\right\} + b(p)\right] = 0,\end{aligned} \tag{7}$$

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 (6) with respect to $\hat{\sigma}$, we have

$$\hat{\sigma} = \left[\frac{\Gamma(3/p)}{\Gamma(1/p)}\right]^{1/2}\left(\frac{p}{N}\sum_{n=0}^{N-1}|y_n|^p\right)^{1/p}. \tag{8}$$

3 Negentropy and Kurtosis

The *entropy* for a continuous complex-valued r.v. Y, which is often called the differential entropy, is defined as

$$H(Y) \triangleq -\int p_Y(v)\log p_Y(v)dv = -\mathcal{E}\left\{\log p_Y(v)\right\}, \tag{9}$$

where $p_Y(.)$ is the pdf of Y. The entropy of a r.v. indicates how much information the observation of the variable provides. Accordingly the large entropy indicates that the variables are really random and contain unstructured information. As mentioned previously, a Gaussian variable has the largest entropy among all r.v.s of equal variance [6].

There are two popular criteria of nongaussianity, namely, negentropy and kurtosis, both of which are frequently used in the field of ICA.

Negentropy J for a complex-valued r.v. Y is defined as

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

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

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

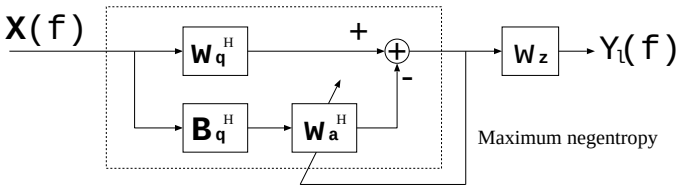


Figure 9: Schematic of a generalized sidelobe canceling (GSC) beamformer for an active source.

where n is the dimension Y. In Section 2, we calculate $H(Y)$ in (10) with a number of super-Gaussian pdfs. Note that negentropy is non-negative, and it is zero if and only if Y has a Gaussian distribution.

The *excess kurtosis* or simply kurtosis of a r.v. Y with zero mean, defined as

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

is another well-known measure of how *non-Gaussian* Y is [6]. 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. This fact manifests itself in Fig. 1, where it is clear that as the kurtosis increases, the pdf becomes more and more spikey and heavy-tailed. Observe that the kurtosis of the GG pdf can be controlled by adjusting the shape parameter p. The detail is explained in Section 5.

Kurtosis can be calculated by simply averaging samples

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

This kurtosis criterion does not require any pdf assumption. Due to its simplicity, it is widely used as a measure of nongaussianity. However, the value of kurtosis might be greatly influenced by a few samples with a low observation probability. Hyvärinen and Oja [6] noted that negentropy was 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.

4 Beamforming and Post-Filtering

Consider a subband beamformer in the GSC configuration [4, §6.7.3] with a post-filter, as shown in Fig. 9. The output of a beamformer for a given subband can be expressed as

$$Y = (\mathbf{w}_\text{q} - \mathbf{B}\mathbf{w}_\text{a})^H \mathbf{X}, \tag{13}$$

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

In keeping with the GSC formalism, $\mathbf{w}_\text{q}$ is chosen to give unity gain in the desired *look direction* [4, §6.7.3]; i.e., to satisfy a *distortionless constraint*. 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 $\mathbf{w}_\text{a}$. While the active weight vector $\mathbf{w}_\text{a}$ is typically chosen to maximize the signal-to-noise ratio (SNR), here we will develop an optimization procedure to find that $\mathbf{w}_\text{a}$ which *maximizes* the negentropy $J(Y)$ described in Section 3.

In order to calculate the negentropy, the variance of the output Y is needed. Substituting (13) into the definition $\sigma_Y^2 = \mathcal{E}\{Y\,Y^*\}$ of variance, we find

$$\begin{aligned}\sigma_Y^2 &= \mathcal{E}\left\{(\mathbf{w}_\text{q} - \mathbf{B}\mathbf{w}_\text{a})^H \mathbf{X}\mathbf{X}^H (\mathbf{w}_\text{q} - \mathbf{B}\mathbf{w}_\text{a})\right\} \\ &= (\mathbf{w}_\text{q} - \mathbf{B}\mathbf{w}_\text{a})^H \boldsymbol{\Sigma}_\mathbf{X} (\mathbf{w}_\text{q} - \mathbf{B}\mathbf{w}_\text{a}),\end{aligned} \tag{14}$$

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

Maximizing the negentropy criterion yields a weight vector $\mathbf{w}_\mathrm{a}$ capable of canceling interferences including incoherent noise that leaks through the sidelobes without the signal cancellation problems encountered in conventional beamforming.

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

$$w_z = \frac{\frac{2}{M(M-1)} \left| \sum_{k=1}^{M-1} \sum_{l=k+1}^{M} \hat{\phi}_{kl} \right|}{\frac{1}{M} \sum_{k=1}^{M} \hat{\phi}_{kk}} \tag{15}$$

where $\hat{\phi}_{kk}$ is the auto-spectral density of the time-aligned input at microphone k and $\hat{\phi}_{kl}$ is the cross-spectral density (CSD) at microphone k and l. The estimation of a desired signal can be improved by averaging the CSDs [9]. The final output of the beamformer and post-filter combination is

$$Y_\mathrm{f} = w_z Y = w_z \left(\mathbf{w}_\mathrm{q} - \mathbf{B}\mathbf{w}_\mathrm{a}\right)^H \mathbf{X}. \tag{16}$$

For the experiments described in Section 7, subband analysis and synthesis were performed with a uniform DFT filter bank based on the modulation of a single prototype impulse response [16], 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, a *regularization* term is often applied that penalizes large active weights, and thereby improves robustness by inhibiting the formation of excessively large sidelobes [4, §6.10]. Such a regularization term can be applied in the present instance by defining the modified optimization criterion

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

for some real $\alpha > 0$.

4.1 Estimation of Active Weights under the Γ pdf

Here we describe necessary formulae for estimating the active weight vectors in the case that the Γ pdf assumption is used.

Substituting (1) and (11) into (10), we can express the negentropy as

$$J(Y) = \log\left|\sigma_Y^2\right| + n\left(1 + \log 2\pi\right) + \frac{1}{T}\sum_{t=0}^{T-1} \log p_Y(Y_t). \tag{18}$$

We maximize the objective function which is the sum of the negentropy and the regularization term. In the absence of a closed-form solution for the $\mathbf{w}_\mathrm{a}$ maximizing the negentropy (18), we must use a numerical optimization algorithm. Such an optimization algorithm typically requires gradient information.

By substituting (18) into (17) and taking the partial derivative on both sides, we obtain the gradient function,

$$\begin{aligned} \frac{\partial \mathcal{J}(Y;\alpha)}{\partial {\mathbf{w}_\mathrm{a}}^*} &= \frac{\partial J(Y;\alpha)}{\partial {\mathbf{w}_\mathrm{a}}^*} + \alpha \mathbf{w}_\mathrm{a} \\ &= \frac{1}{|\sigma_Y^2|}\frac{\partial |\sigma_Y^2|}{\partial {\mathbf{w}_\mathrm{a}}^*} + \frac{1}{T}\sum_{t=0}^{T-1} \frac{1}{p_Y(Y_t)} \frac{\partial p_Y(Y_t)}{\partial {\mathbf{w}_\mathrm{a}}^*} + \alpha \mathbf{w}_\mathrm{a} \end{aligned} \tag{19}$$

where

$$\frac{\partial |\sigma_Y^2|}{\partial {\mathbf{w}_\mathrm{a}}^*} = \mathcal{E}\left\{-\mathbf{B}^H \mathbf{X} Y^*\right\}. \tag{20}$$

Equations (19) and (20) are sufficient to implement a numerical optimization algorithm based, for example, on the method of *conjugate gradients* [17, §1.6], whereby the negentropy $J(Y)$ can be maximized.

4.2 Estimation of Active Weights under the Generalized Gaussian pdf

4.2.1 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 beamformer's output as the r.v. for calculating the entropy. By substituting (4) and (11) into (10), we have the negentropy

$$J(Y) = \log\left|\sigma_Y^2\right| + n\,(1+\log 2\pi) - H_{\mathrm{GG}}(|Y|). \tag{21}$$

In order to apply the gradient algorithm, we derive the gradient information again. By substituting (21) into (17) and taking the partial derivative on both sides, where the shape parameter is fixed, we can obtain the objective function

$$\frac{\partial \mathcal{J}(Y;\alpha)}{\partial \mathbf{w}_{\mathrm{a}}{}^*} = \frac{1}{\sigma_Y^2}\frac{\partial \sigma_Y^2}{\partial \mathbf{w}_{\mathrm{a}}{}^*} - \frac{\partial H_{\mathrm{GG}}(|Y|)}{\partial \mathbf{w}_{\mathrm{a}}{}^*} + \alpha \mathbf{w}_{\mathrm{a}} \tag{22}$$

where

$$\frac{\partial H_{\mathrm{GG}}(|Y|)}{\partial \mathbf{w}_{\mathrm{a}}{}^*} = \frac{1}{\hat{\sigma}_{|Y|}}\frac{\partial \hat{\sigma}_{|Y|}}{\partial \mathbf{w}_{\mathrm{a}}{}^*}. \tag{23}$$

Taking the derivative on both sides of (8), we find

$$\begin{aligned}\frac{\partial \hat{\sigma}_{|Y|}}{\partial \mathbf{w}_{\mathrm{a}}{}^*} = \frac{p}{T}\left[\frac{\Gamma(3/p)}{\Gamma(1/p)}\right]^{\frac{1}{2}} &\times \left[\frac{p}{T}\sum_{t=0}^{T-1}|Y_t|^p\right]^{\frac{1}{p}-1} \\ &\times \left[\sum_{t=0}^{T-1}|Y_t|^{p-1}\frac{\partial |Y_t|}{\partial \mathbf{w}_{\mathrm{a}}{}^*}\right],\end{aligned} \tag{24}$$

where the gradient of the magnitude is

$$\frac{\partial |Y_t|}{\partial \mathbf{w}_{\mathrm{a}}{}^*} = -\frac{1}{2|Y_t|}\mathbf{B}^H\mathbf{X}Y_t^*. \tag{25}$$

We can implement a numerical optimization algorithm from equations (22) to (25).

4.2.2 Parameter optimization 2

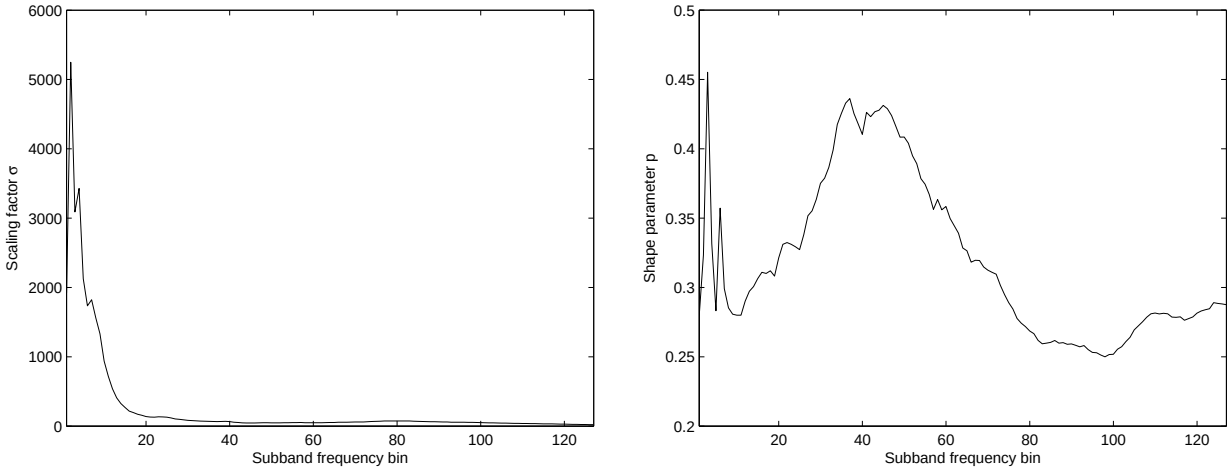
One might think that the entropy of the GG pdf for the complex valued r.v. could be approximated by assuming that real and imaginary parts are independent. With such an assumption, we can express the differential entropy of the GG pdf as

$$H(Y) \approx H_r(Y_r) + H_i(Y_i) \tag{26}$$

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 (26) into (10) and adding the regularization term, we obtain the objective function

$$\begin{aligned}\mathcal{J}(Y;\alpha) = &\log\left|\sigma_Y^2\right| + n\,(1+\log 2\pi) \\ &- H_r(Y_r) - H_i(Y_i) + \alpha\|\mathbf{w}_{\mathrm{a}}\|^2.\end{aligned} \tag{27}$$



(a) (b)

Figure 10: The parameters of the GG pdf for each frequency bin; (a) scaling parameter $\hat{\sigma}_{|Y|}$ and (b) shape parameter p, where the sampling frequency is 16 kHz.

In order to employ the gradient algorithm, we take the partial derivative of (27)

$$\frac{\partial \mathcal{J}(Y;\alpha)}{\partial \mathbf{w}_{\mathrm{a}}{}^{*}} = \frac{1}{|\sigma_Y^2|}\frac{\partial |\sigma_Y^2|}{\partial \mathbf{w}_{\mathrm{a}}{}^{*}} - \frac{\partial H_r(Y_r)}{\partial \mathbf{w}_{\mathrm{a}}{}^{*}} - \frac{\partial H_i(Y_i)}{\partial \mathbf{w}_{\mathrm{a}}{}^{*}} + \alpha \mathbf{w}_{\mathrm{a}}, \tag{28}$$

where

$$\frac{\partial |Y_{r,t}|}{\partial \mathbf{w}_{\mathrm{a}}{}^{*}} = -\frac{1}{2}\mathbf{B}^{H}\mathbf{X} \cdot \mathrm{sign}(Y_{r,t}) \tag{29}$$

and

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

Equations (28) through (30) are sufficient for implementing the gradient algorithm.

5 Speech Modeling with the GG pdf

Subbands of speech can be precisely modeled by estimating the parameters of the GG pdf from training samples. The trained parameters give us an intuitive insight into the speech pdf. Fig. 10 shows the scaling parameter $\hat{\sigma}_{|Y|}$ and the shape parameter p for each frequency bin. The training samples used for estimating the GG pdf here were taken from clean speech data in the SSC development set [1].

It is clear from Fig. 10 that the scaling parameter $\hat{\sigma}_{|Y|}$ becomes smaller at higher frequency. We can consider that the scaling parameter $\hat{\sigma}_{|Y|}$ can be associated with the variance. Therefore Fig. 10 implies that the magnitude at lower frequency varies more than that at higher frequency. Notice that $\hat{\sigma}_{|Y|}$ doesn't indicate the variance exactly in the case that the ML method is used for estimating it. We can see from Fig. 10 that the GG pdfs trained with actual speech data are super-Gaussian $p < 2$ for all subbands. Moreover, the subband samples of speech are more super-Gaussian than the Laplace pdf given that $p < 1$ for all frequency bins.

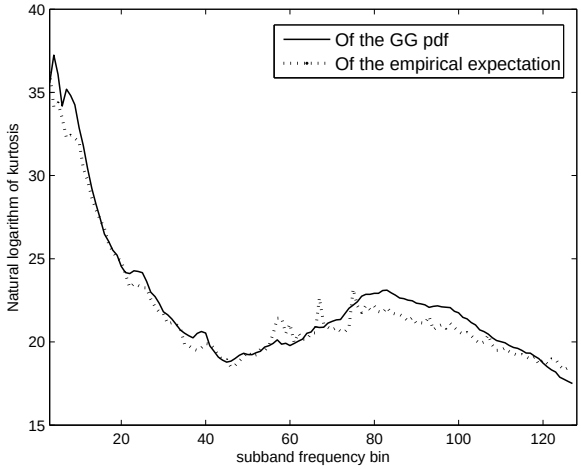


Figure 11: Kurtosis for each frequency bin, where the sampling frequency is 16 kHz.

As mentioned previously, the kurtosis is a measure of the super-Gaussianity of a pdf. Therefore it might be interesting to see the kurtosis of the GG pdf. The kurtosis of a GG pdf can be expressed as

$$\mathrm{kurt}(Y_{gg}) = \hat{\sigma}^4 \left\{ \frac{\Gamma(1/p)\Gamma(5/p)}{\Gamma(3/p)^2} - 3 \right\}. \tag{31}$$

A derivation of (31) is provided in the Appendix. Fig. 11 shows kurtosis values for all frequency bins. In Fig. 11, a solid line indicates the kurtosis of the GG pdf calculated with Eq. (31) and a broken line presents the empirical kurtosis computed with Eq. (12). It is clear from Fig. 11 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. 11 that kurtosis becomes smaller at higher frequency, which indicates that the pdf of lower frequency components are more super-Gaussian than those of higher frequency ones.

6 Simulation

Conventional adaptive beamforming algorithms determine the optimum weight vector that minimizes the beamformer's output:

$$\mathbf{w}^H \mathbf{\Sigma_X} \mathbf{w}, \tag{32}$$

subject to the distortionless constraint for the desired 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 [18]. The weight vector of the MVDR beamformer can be expressed as

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

A small value is typically added to the diagonal of $\mathbf{\Sigma_X}$ in order to ensure that the matrix is invertible. The conventional beamformers as well as the MVDR beamfomer would attempt to null out any interfering signal. However, it leads to the signal cancellation problem [5] in the case that there is an interference signal which is correlated with a desired signal. In realistic environments, interference

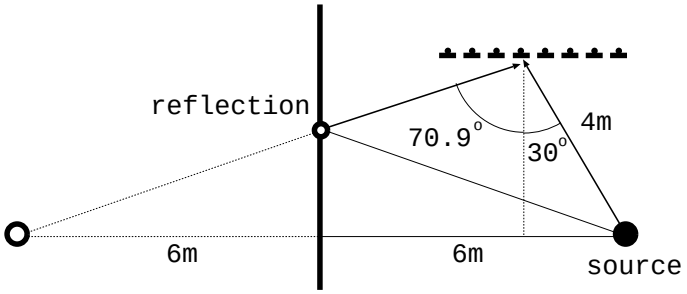


Figure 12: Configuration of a source, sensors, and reflective surface for simulation.

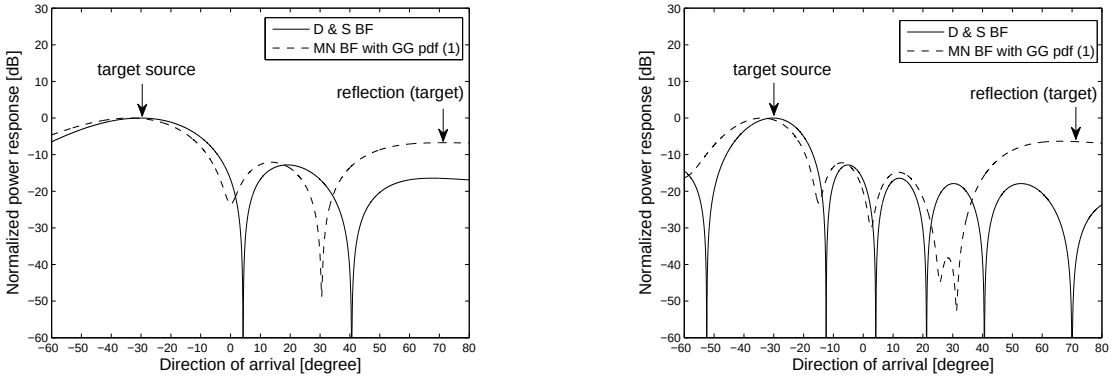
signals are highly correlated with a target signal since the target signal is reflected from hard surfaces such as walls and tables. Therefore, the adaptation of the weight vector is usually halted whenever the desired source is active. Although many techniques have been proposed to avoid the signal cancellation, one of the most successful beamforming algorithms is perhaps the robust beamformer with the GSC configuration proposed by Hoshuyama *et al.* [19]. 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. Then their algorithm is not able to suppress the leaked interference signals. In reality, the interference signals often come in the upper branch because of the steering error and spatial aliasing. Therefore it could be considered that Hoshuyama's algorithm has a trade-off problem between the avoidance of the signal cancellation and suppression of the interference signals. Such a trade-off problem can be solved by simply halting the adaptation of the ABM and only update the active weight vectors in the case of a high SNR [20]. However, such a switching algorithm based on the SNR requires complicated rules which are generally determined empirically.

Conventional robust beamforming algorithms fundamentally have tackled how to remove 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 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, and could thus be removed through a suitable choice of $\mathbf{w}_{\mathrm{a}}$. 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.

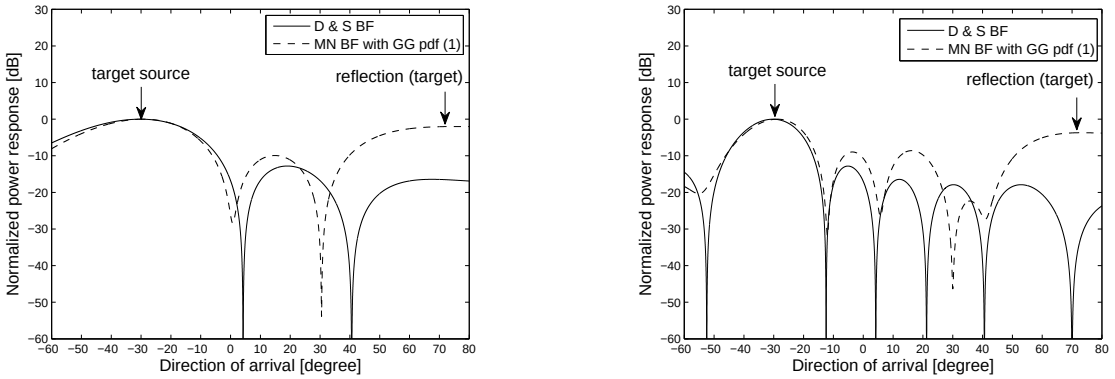
In order to verify that the MN beamforming algorithm forms sidelobes directed towards the reflection of a desired signal, we conducted experiments with a simulated acoustic environment. As shown in Fig. 12, we considered a simple configuration with a sound source, a reflective surface, and a linear array of eight microphones positioned with 10 cm intersensor spacing. Actual speech data were used as a source in this simulation, which was based on the *image method* [21]. Fig. 13 shows beam patterns at $f_{\mathrm{s}} = 800$ Hz and $f_{\mathrm{s}} = 1500$ Hz obtained with delay-and-sum (D&S) beamformer and the MN beamforming algorithm with the GG pdf of the magnitude.

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. Once more, it is apparent that the MN beamformer emphasizes the



(a) (b)

Figure 13: Beam patterns produced by the delay-and-sum beamformer and the MN beamforming algorithm using a spherical wave assumption for (a) $f_{\mathrm{s}} = 800$ Hz and (b) $f_{\mathrm{s}} = 1500$ Hz.



(a) (b)

Figure 14: Beam patterns produced by the delay-and-sum beamformer and the MN beamforming algorithm using a plane wave assumption for (a) $f_{\mathrm{s}} = 800$ Hz and (b) $f_{\mathrm{s}} = 1500$ Hz.

reflection from the desired source.

From the statistical point of view, the difference between the conventional and MN beamformers is that the MN beamforming algorithm takes account into the high-order statistics (HOS). On the other hand, the conventional beamformers are based only on the consideration of the covariance, the second-order statistics (SOS). Therefore the simulation results suggest that the measure of the HOS could be associated with how much a target signal is enhanced with its reflections.

7 Experiments

We performed far-field automatic speech recognition (ASR) experiments on the *Multi-Channel Wall Street Journal Audio Visual Corpus* (MC-WSJ-AV) from the *Augmented Multi-party Interaction* (AMI); see Lincoln *et al.* [1] for a description of the data collection apparatus. In the *single speaker stationary* scenario of the MC-WSJ-AV, a speaker was asked to sit or stand in front of a presentation screen and read sentences from different positions. The far-field speech data was recorded with two circular, eight-channel microphone arrays in a reverberant room. In addition to reverberation, some

recordings include significant amounts of background noise. The sampling rate of the recordings was 16 kHz. As the data was recorded with real speakers in a realistic acoustic environment and not artificially convolved with measured room impulse responses, the positions of the speakers' heads as well as the speaking volume vary even though the speakers are largely stationary. Indeed, it is exactly this behavior of real speakers that makes working with data from corpora such as MC-WSJ-AV so much more challenging than working with data that was played through a loud speaker into a room, not to mention data that was *artificially convolved*.

Our test data set for the experiments contains recordings of 10 speakers where each speaker reads approximately 40 sentences taken from the 5,000 word vocabulary Wall Street Journal (WSJ) task. It gives a total of 352 utterances which correspond to 39.2 minutes of speech. There are a total of 11,598 word tokens in the reference transcriptions. The test data do not include training data.

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 a half wavelength, thereby causing *spatial aliasing* [4, §2.5].

Prior to beamforming, we first estimated the speaker's position with the *Orion* source tracking system [22]. 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 [9] was performed after beamforming. The feature extraction of our ASR system was based on cepstral features estimated with a warped *minimum variance distortionless response* [23] (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 [24].

The far-field ASR experiments reported here were conducted with a *word trace decoder* implemented along the lines suggested by Saon *et al.* [25]. The decoder is capable of generating word lattices, which can then be optimized with weighted finite-state transducer (WFST) operations as in [26]; 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 [27].

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 [28, §12], and speaker-adapted training under a ML criterion (ML-SAT) [29]. 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 [30]. A description of the four decoding passes follows:

Table 1: 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
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

1. Decode with the unadapted, conventional ML acoustic model and bigram language model (LM).
2. Estimate vocal tract length normalization (VTLN) [31] parameters and constrained maximum likelihood linear regression parameters (CMLLR) [32] for each speaker, then redecode with the conventional ML acoustic model and bigram LM.
3. Estimate VTLN, CMLLR, and maximum likelihood linear regression (MLLR) [33] parameters for each speaker, then redecode with the conventional model and bigram LM.
4. Estimate VTLN, CMLLR, MLLR parameters for each speaker, then redecode with the ML-SAT model and bigram LM.

Table 1 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 described in Table 1. It is clear from Table 1 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 1 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 because the GG pdf can model the magnitude of the subband of speech best by training the shape parameter at each subband frequency bin. 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 inpendent (MN BF with GG pdf (2)). We found it better to treat the subband components as spherically-invariant random processes (SIRPs) as in [2, 12] and are led to conclude that the real and imaginary parts are dependent as mentioned in [8]. Table 1 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 also known as the minimum mean-squared-error beamformer (MMSE BF) [18, §3]. One can see from Table. 1 that the MVDR beamformer with post-filtering (MMSE BF) provides better recognition performance than D&S BF with PF. Notice MVDR beamforming algorithms require speech activity detection in order to avoid the signal cancellation. For the adaptation of the MVDR beamformer, we used the first 0.1 and last 0.1 seconds in each utterance data which contain only background noise. Again, in contrast to conventional beamforming methods, our algorithm doesn't 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.

We also examine the effect of the regularization expressed in (17). Table 2 shows the WERs against the regularization parameter α, where we used the MN beamforming algorithm with the GG

Table 2: 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

pdf of the magnitude r.v.. We can see from Table 2 that the regularization parameter $\alpha = 10^{-2}$ provides the best result although its impact on the recognition performance is not significant. 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 reliable enough to completely determine the active weight vector due to the steering error.

8 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 adaptive 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 the far-field speech recognition.

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

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

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

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

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

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

By substituting the 2nd and 4th moments obtained from Equation (37), the kurtosis of the GG pdf $\text{kurt}\,(Y_{gg})$ can now be expressed as

$$\frac{A(p,\hat{\sigma})^4}{p\Gamma(1+1/p)}\Gamma\left(5/p\right) - 3\left\{\frac{A(p,\hat{\sigma})^2}{p\Gamma(1+1/p)}\Gamma\left(3/p\right)\right\}^2. \tag{38}$$

Since the Γ function satisfies $p\Gamma(1+1/p) = \Gamma(1/p)$, equation (38) can be simplified as

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

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)*, Cancun, Mexico, 2005, pp. 357–362.

[2] Kenichi Kumatani, Tobias Gehrig, Uwe Mayer, Emilian Stoimenov, John McDonough, and Matthias Wölfel, "Adaptive beamforming with a minimum mutual information criterion," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 2527–2541, 2007.

[3] Hari Krishna Maganti, Daniel Gatica-Perez, and Iain 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] H. L. Van Trees, *Optimum Array Processing*, Wiley-Interscience, New York, 2002.

[5] Bernard Widrow, Kenneth M. Duvall, Richard P. Gooch, and William 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] Aapo Hyvärinen and Erkki Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, pp. 411–430, 2000.

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

[8] Jan S. Erkelens, Richard C. Hendriks, Richard Heusdens, and Jesp er 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.

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

[10] Robert G. Gallager, *Information Theory and Reliable Communication*, John Wiley & Sons, New York, 1968.

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

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

[13] Stephen Wolfram, *The Mathematica Book*, Cambridge University Press, Cambridge, 3 edition, 1996.

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

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

[16] Kenichi Kumatani, John McDonough, Stefan Schacht, Dietrich Klakow, Philip N. Garner, and Weifeng Li, "Filter bank design for subband adaptive beamforming and application to speech recognition," *Submitted to IEEE Transactions on Signal Processing*, 2008.

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

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

[19] Osamu Hoshuyama, Akihiko Sugiyama, and Akihiro 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.

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

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

[22] Tobias Gehrig, Ulrich Klee, John McDonough, Shajith Ikbal, Matthias Wölfel, and Christian Fügen, "Tracking and beamforming for multiple simultaneous speakers with probabilistic data association filters," in *Proc. Interspeech*, Pittsburgh, Pennsylvania, U.S.A, 2006, pp. 2594–2597.

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

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

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

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

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

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

[29] T. Anastasakos, J. McDonough, R. Schwarz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, Pennsylvania, USA, 1996, pp. 1137–1140.

[30] L. Uebel and P. Woodland, "Improvements in linear transform based speaker adaptation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, U.S.A, 2001.

[31] L. Welling, H. Ney, and S. Kanthak, "Speaker adaptive modeling by vocal tract normalizatio," *IEEE Trans. Speech Audio Proc.*, vol. 10, no. 6, pp. 415–426, 2002.

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

[33] 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.