IDIAP RESEARCH REPORT

Adaptive Beamforming with a Minimum Mutual Information Criterion

Kenichi Kumatani [a] Uwe Mayer [k]
Tobias Gehrig [k] Emilian Stoimenov [k]
John McDonough [b] Matthias Wölfel [k]

IDIAP–RR 07-74

January 2008

[a] IDIAP Research Institute, Martigny, Switzerland
[b] Spoken Language Systems at Saarland University in Saarbrücken, Germany
[k] Institute for Theoretical Computer Science at University of Karlsruhe in Karlsruhe, 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 07-74

Adaptive Beamforming with a Minimum Mutual Information Criterion

Kenichi Kumatani   Uwe Mayer   Tobias Gehrig   Emilian Stoimenov
John McDonough   Matthias Wölfel

January 2008

Abstract. In this work, we consider an acoustic beamforming application where two speakers are simultaneously active. We construct one subband-domain beamformer in *generalized sidelobe canceller* (GSC) configuration for each source. In contrast to normal practice, we then jointly optimize the *active weight vectors* of both GSCs to obtain two output signals with *minimum mutual information* (MMI). Assuming that the subband snapshots are Gaussian-distributed, this MMI criterion reduces to the requirement that the *cross-correlation coefficient* of the subband outputs of the two GSCs vanishes. We also compare separation performance under the Gaussian assumption with that obtained from several super-Gaussian probability density functions (pdfs), namely, the Laplace, K_0, and Γ pdfs. Our proposed technique provides effective nulling of the undesired source, but without the signal cancellation problems seen in conventional beamforming. Moreover, our technique does not suffer from the source permutation and scaling ambiguities encountered in conventional blind source separation algorithms. We demonstrate the effectiveness of our proposed technique through a series of far-field automatic speech recognition experiments on data from the *PASCAL Speech Separation Challenge* (SSC). On the SSC development data, the simple delay-and-sum beamformer achieves a word error rate (WER) of 70.4%. The MMI beamformer under a Gaussian assumption achieves a 55.2% WER, which is further reduced to 52.0% with a K_0 pdf, whereas the WER for data recorded with a close-talking microphone is 21.6%.

1 Introduction

In acoustic beamforming, it is typically assumed that the position of the desired source or speaker as well as the positions of the individual sensors in the array are known. Indeed, knowledge of these positions is what distinguishes beamforming from *blind source separation* (BSS), and, as described below, has the practical effect of eliminating the ambiguities encountered in the latter. A conventional beamformer in *generalized sidelobe canceller* (GSC) configuration is structured such that the direct signal from the speaker is undistorted [1, §6.7.3]. Subject to this *distortionless constraint*, the total output power of the beamformer is minimized through the appropriate adjustment of an *active weight vector*, which effectively places a null on any source of interference, but can also lead to undesireable *signal cancellation*. 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 an acoustic beamforming application where two speakers are simultaneously active. We construct one subband-domain beamformer in GSC configuration for each source. In contrast to normal practice, we then jointly adjust the active weight vectors of both GSCs to obtain two output signals with *minimum mutual information* (MMI). Assuming that the subband snapshots are Gaussian-distributed, this MMI criterion reduces to the requirement that the *cross-correlation coefficient* of the subband outputs of the two GSCs vanishes. Parra and Alvino [2] proposed a *geometric source separation* (GSS) algorithm with many similarities to the algorithm proposed here. Their algorithm attempts to decorrelate the outputs of two beamformers. In Section 3.3, we discuss Parra and Alvino's GSS algorithm, and the experimental results presented in Section 5 demonstrate that our algorithm provides superior separation performance.

We demonstrate that our proposed technique provides effective nulling of the undesired source, but without the signal cancellation problems seen in conventional beamforming. This allows the adaptation of the active weight vectors to continue when *both* sources are active. Moreover, our technique does not suffer from the source permutation and scaling ambiguities encountered in conventional frequency- and subband-domain BSS algorithms [3]. In addition to the Gaussian assumption that is commonly made in conventional beamforming, we also investigate several super-Gaussian densities that are more commonly used in the field of *independent component analysis* [4]. We demonstrate the effectiveness of our proposed technique through a series of far-field automatic speech recognition experiments on data from the *PASCAL Speech Separation Challenge* [5].

The balance of this work is organized as follows. In Section 2, we review the definition of mutual information and demonstrate that, under a Gaussian assumption, the mutual information of two complex random variables is a simple function of their cross-correlation coefficient. We discuss our MMI beamforming criterion in Section 3, and compare it to the decorrelation approach of Parra and Alvino [2]. Section 4 presents the framework needed to apply minimum mutual information beamforming when the Gaussian assumption is relaxed. In particular, we develop multivariate pdfs for the Laplace, K_0, and Γ density functions, and then develop parameter estimation formulae based on these for optimizing the active weight vector of a GSC. In Section 5, we present the results of far-field automatic speech recognition experiments. Finally, in Section 6, we present our conclusions and plans for future work. The appendix presents a derivation of the multivariate pdfs for the Laplace, K_0, and Γ densities based on the *Meijer G-function* [6].

2 Mutual Information

Here we derive the mutual information of two zero-mean Gaussian random variables (r.v.s). Consider two r.v.s Y_1 and Y_2. By definition, the *mutual information* [7] of Y_1 and Y_2 is given by

$$\begin{aligned} I(Y_1, Y_2) &= \mathcal{E}\left\{\log \frac{p(Y_1, Y_2)}{p(Y_1)p(Y_2)}\right\} \\ &= \mathcal{E}\{\log p(Y_1, Y_2)\} - \mathcal{E}\{\log p(Y_1)\} - \mathcal{E}\{\log p(Y_2)\} \end{aligned} \tag{1}$$

where $\mathcal{E}\{\}$ denotes ensemble expectation. For Gaussian r.v.s, we have

$$p(y_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-y_i^2/2\sigma_i^2}.$$

Hence, we can solve the latter two expectations in (1) as

$$\begin{aligned}\mathcal{E}\{\log p(Y_i)\} &= \mathcal{E}_{Y_i}\left\{-\frac{1}{2}\log 2\pi\sigma_i^2 - \frac{1}{2}\frac{Y_i^2}{\sigma_i^2}\right\} \\ &= -\frac{1}{2}\log 2\pi\sigma_i^2 - \frac{1}{2}\int_{-\infty}^{\infty}\frac{y_i^2}{\sigma_i^2}\,p(y_i)\,dy_i. \end{aligned} \tag{2}$$

For jointly Gaussian r.v.s,

$$p(Y_1, Y_2) = \frac{1}{\sqrt{|2\pi\Sigma|}}\exp\left[-\frac{1}{2}\mathbf{Y}^T\Sigma^{-1}\mathbf{Y}\right]$$

where $\mathbf{Y} = \begin{bmatrix} Y_1 & Y_2 \end{bmatrix}^T$ and the *covariance matrix* of $\mathbf{Y}$ is given by [8, §2.3]

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} \\ \sigma_1\sigma_2\rho_{12} & \sigma_2^2 \end{bmatrix} \tag{3}$$

with

$$\rho_{12} = \frac{\epsilon_{12}}{\sigma_1\,\sigma_2}$$

where $\epsilon_{12} = \mathcal{E}\{Y_1\,Y_2^*\}$. Hence, the first expectation in (1) can be rewritten as

$$\begin{aligned}\mathcal{E}\{\log p(Y_1, Y_2)\} &= \mathcal{E}\left\{-\frac{1}{2}\log|2\pi\Sigma| - \frac{1}{2}\mathbf{Y}^T\Sigma^{-1}\mathbf{Y}\right\} \\ &= -\frac{1}{2}\log|2\pi\Sigma| - \frac{1}{2}\int_{\mathbf{Y}}\mathbf{Y}^T\Sigma^{-1}\mathbf{Y}\,p(\mathbf{Y})\,d\mathbf{Y}. \end{aligned} \tag{4}$$

Due to the *whitening* [9, §2.3] provided by the term Σ^{-1}, the integral in (4) decouples into two integrals of the form of the integral in (2). Hence, when (2) and (4) are substituted back into (1), the integral terms cancel out, and what remains is

$$\begin{aligned}I(Y_1, Y_2) = &-\tfrac{1}{2}\log\left[4\pi^2\sigma_1^2\sigma_2^2(1-\rho_{12}^2)\right] \\ &+\tfrac{1}{2}\log 2\pi\sigma_1^2 + \tfrac{1}{2}\log 2\pi\sigma_2^2\end{aligned}$$

or, upon cancelling common terms,

$$I(Y_1, Y_2) = -\tfrac{1}{2}\log\left(1-\rho_{12}^2\right).$$

For the complex r.v.s considered in the next section, it is straightforward to show that

$$I(Y_1, Y_2) = -\log\left(1 - |\rho_{12}|^2\right). \tag{5}$$

From (5) it is clear that minimizing the mutual information between two zero-mean Gaussian r.v.s is equivalent to minimizing the squared magnitude of their cross correlation coefficient and that

$$I(Y_1, Y_2) = 0 \leftrightarrow |\rho_{12}| = 0.$$

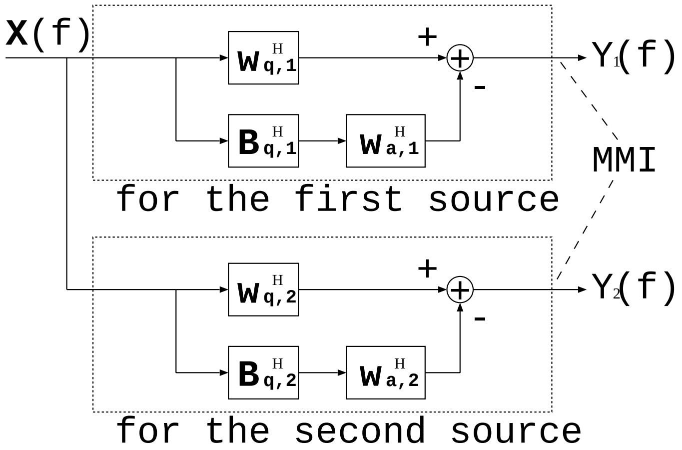


Figure 1: Schematic of generalized sidelobe cancelling (GSC) beamformers for each active source.

3 Beamforming

Consider a subband beamformer in GSC configuration [1, §6.7.3]. Assuming there are two such beamformers aimed at different sources, as shown in Fig. 1, the output of the ith beamformer for a given subband can be expressed as,

$$Y_i = \left(\mathbf{w}_{\mathrm{q},i} - \mathbf{B}_i \mathbf{w}_{\mathrm{a},i}\right)^H \mathbf{X} \tag{6}$$

where $\mathbf{w}_{\mathrm{q},i}$ is the *quiescent weight vector* for the ith source, $\mathbf{B}_i$ is the *blocking matrix*, $\mathbf{w}_{\mathrm{a},i}$ is the *active weight vector*, and $\mathbf{X}$ is the input subband *snapshot vector*, which is common to both sources. In keeping with the GSC formalism, $\mathbf{w}_{\mathrm{q},i}$ is chosen to give unity gain in the desired *look direction* [1, §6.7.3]; i.e., to satisfy a *distortionless constraint*. The blocking matrix $\mathbf{B}_i$ is chosen to be orthogonal to $\mathbf{w}_{\mathrm{q},i}$, such that

$$\mathbf{B}_i^H \, \mathbf{w}_{\mathrm{q},i} = \mathbf{0}.$$

This orthogonality implies that the distortionless constraint will be satisfied for any $\mathbf{w}_{\mathrm{a},i}$. While the active weight vector $\mathbf{w}_{\mathrm{a},i}$ is typically chosen to maximize the signal-to-noise ratio (SNR), here we will develop an optimization procedure to find the $\mathbf{w}_{\mathrm{a},i}$ that *minimizes* the mutual information $I(Y_1, Y_2)$; based on the development of Section 2, this implies minimizing $|\rho_{12}|$. Upon substituting (6), the variance $\sigma_i^2 = \mathcal{E}\left\{Y_i \, Y_i^*\right\}$ of Y_i can be expressed as

$$\sigma_i^2 = \left(\mathbf{w}_{\mathrm{q},i} - \mathbf{B}_i \mathbf{w}_{\mathrm{a},i}\right)^H \Sigma_{\mathbf{X}} \left(\mathbf{w}_{\mathrm{q},i} - \mathbf{B}_i \mathbf{w}_{\mathrm{a},i}\right) \tag{7}$$

where $\Sigma_{\mathbf{X}} = \mathcal{E}\{\mathbf{X}\mathbf{X}^H\}$ is the covariance matrix of the snapshot $\mathbf{X}$. The cross-correlation coefficient ρ_{12} between Y_1 and Y_2 can be expressed as [8, §2.3]

$$\rho_{12} = \frac{\epsilon_{12}}{\sigma_1 \, \sigma_2} \tag{8}$$

where

$$\begin{aligned} \epsilon_{12} &= \mathcal{E}\{Y_1 \, Y_2^*\} \\ &= \left(\mathbf{w}_{\mathrm{q},1} - \mathbf{B}_1 \mathbf{w}_{\mathrm{a},1}\right)^H \Sigma_{\mathbf{X}} \left(\mathbf{w}_{\mathrm{q},2} - \mathbf{B}_2 \mathbf{w}_{\mathrm{a},2}\right). \end{aligned} \tag{9}$$

Hence,

$$|\rho_{12}|^2 = \frac{|\epsilon_{12}|^2}{\sigma_1^2 \, \sigma_2^2}. \tag{10}$$

Minimizing the mutual information criterion yields a weight vector $\mathbf{w}_{\mathrm{a},i}$ capable of canceling interference that leaks through the sidelobes without the signal cancellation problems encountered in conventional beamforming.

For the experiments described in Section 5, subband analysis and resynthesis were performed with a DFT filter bank based on the modulation of a single prototype impulse response [10, §8], which was designed to achieve *perfect reconstruction* (PR). 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 savings with respect to a time-domain filter-and-sum beamformer with filters of the same length on the output of each sensor. In addition, the GSC constraint imposed here resolves problems of source permutation and scaling ambiguity encountered in conventional frequency- and subband-domain BSS algorithms [3]. Although the PR filter bank achieves good separation, as demonstrated by the results in Section 5, de Haan *et al.* [11] point out that the PR design may in fact be suboptimal for applications involving beamforming or adaptive filtering. This point is discussed briefly in Section 6.

3.1 Parameter Optimization

In the absence of a closed-form solution for those $\mathbf{w}_{\mathrm{a},i}$ minimizing $|\rho_{12}|^2$, we must use a numerical optimization algorithm. Such an optimization algorithm typically requires gradient information. Hence, let us apply the chain rule [1, §A.7.4] to (10), and write

$$\begin{aligned}\frac{\partial|\rho_{12}|^2}{\partial\mathbf{w}^*_{\mathrm{a},1}} &= \frac{1}{\sigma_1^4\,\sigma_2^4}\left(\frac{\partial\epsilon_{12}}{\partial\mathbf{w}^*_{\mathrm{a},1}}\epsilon^*_{12}\sigma_1^2\sigma_2^2 - \frac{\partial\sigma_1^2}{\partial\mathbf{w}^*_{\mathrm{a},1}}|\epsilon_{12}|^2\sigma_2^2\right)\\ &= \frac{1}{\sigma_1^4\,\sigma_2^4}\Big[-\mathbf{B}_1^H\Sigma_{\mathbf{X}}(\mathbf{w}_{\mathrm{q},2}-\mathbf{B}_2\mathbf{w}_{\mathrm{a},2})\epsilon^*_{12}\sigma_1^2\sigma_2^2\\ &\qquad +\mathbf{B}_1^H\Sigma_{\mathbf{X}}(\mathbf{w}_{\mathrm{q},1}-\mathbf{B}_1\mathbf{w}_{\mathrm{a},1})|\epsilon_{12}|^2\sigma_2^2\Big].\end{aligned}$$

The last equation can be simplified to

$$\begin{aligned}\frac{\partial|\rho_{12}|^2}{\partial\mathbf{w}^*_{\mathrm{a},1}} = \frac{1}{\sigma_1^4\,\sigma_2^4}\mathbf{B}_1^H\Sigma_{\mathbf{X}}\Big[&|\epsilon_{12}|^2\sigma_2^2(\mathbf{w}_{\mathrm{q},1}-\mathbf{B}_1\mathbf{w}_{\mathrm{a},1})\\ &-\epsilon^*_{12}\sigma_1^2\sigma_2^2(\mathbf{w}_{\mathrm{q},2}-\mathbf{B}_2\mathbf{w}_{\mathrm{a},2})\Big].\end{aligned} \tag{11}$$

From symmetry it then follows

$$\begin{aligned}\frac{\partial|\rho_{12}|^2}{\partial\mathbf{w}^*_{\mathrm{a},2}} = \frac{1}{\sigma_1^4\,\sigma_2^4}\mathbf{B}_2^H\Sigma_{\mathbf{X}}\Big[&|\epsilon_{12}|^2\sigma_1^2(\mathbf{w}_{\mathrm{q},2}-\mathbf{B}_2\mathbf{w}_{\mathrm{a},2})\\ &-\epsilon_{12}\sigma_1^2\sigma_2^2(\mathbf{w}_{\mathrm{q},1}-\mathbf{B}_1\mathbf{w}_{\mathrm{a},1})\Big].\end{aligned} \tag{12}$$

Equations (11) and (12) are sufficient to implement a numerical optimization algorithm based, for example, on the method of *conjugate gradients* [12, §1.6], whereby $|\rho_{12}|^2$ is directly minimized. Alternatively, the mutual information $I(Y_1, Y_2)$ can be minimized, which is the approach we have adopted here, as the minimization of mutual information under a Gaussian assumption can be readily extended to the super-Gaussian pdfs considered in Section 4. To formulate an algorithm for minimizing $I(Y_1, Y_2)$, we need only begin from (5) and write

$$\frac{\partial I(Y_1,Y_2)}{\partial\mathbf{w}^*_{\mathrm{a},i}} = \frac{1}{2\,(1-|\rho_{12}|^2)}\cdot\frac{\partial|\rho_{12}|^2}{\partial\mathbf{w}^*_{\mathrm{a},i}} \tag{13}$$

which, together with (11) and (12), is sufficient to calculate the required gradients.

3.2 Regularization

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 [1, §6.10]. Such a regularization term can be applied in the present instance by defining the modified optimization criterion

$$\mathcal{I}(Y_1, Y_2; \alpha) = I(Y_1, Y_2) + \alpha \|\mathbf{w}_{\mathrm{a},1}\|^2 + \alpha \|\mathbf{w}_{\mathrm{a},2}\|^2 \tag{14}$$

for some real $\alpha > 0$. Taking the partial derivative on both sides of (14) yields

$$\frac{\partial \mathcal{I}(Y_1, Y_2; \alpha)}{\partial \mathbf{w}_{\mathrm{a},i}^*} = \frac{1}{2\,(1 - |\rho_{12}|^2)} \cdot \frac{\partial |\rho_{12}|^2}{\partial \mathbf{w}_{\mathrm{a},i}^*} + \alpha \mathbf{w}_{\mathrm{a},i}. \tag{15}$$

3.3 Geometric Source Separation

Parra and Alvino [2] proposed a *geometric source separation* (GSS) algorithm with many similarities to the algorithm proposed here. Their work was based on two beamformers with geometric constraints that made them functionally equivalent to GSC beamformers. The principal difference between GSS and the algorithm proposed here is that GSS seeks to minimize $|\epsilon_{12}|^2$ instead of $|\rho_{12}|^2$.

Although the difference between minimizing $|\epsilon_{12}|^2$ instead of $|\rho_{12}|^2$ may seem very slight, it can in fact lead to radically different behavior. To achieve the desired optimum, both criteria will seek to place deep nulls on the unwanted source; this characteristic is associated with $|\epsilon_{12}|^2$, which also comprises the *numerator* of $|\rho_{12}|^2$. Such null steering is also observed in conventional adaptive beamformers [1, §6.3]. The difference between the two optimization criteria is due to the presence of the terms σ_i^2 in the denomimnator of $|\rho_{12}|^2$, which indicate that, in addition to nulling out the unwanted signal, an improvement of the objective function is also possible by *increasing* the strength of the desired signal. For acoustic beamforming in realistic environments, there are typically strong reflections from hard surfaces such as tables and walls. A conventional beamformer would attempt to null out strong reflections of an interfering signal, but strong reflections of the desired signal can lead to signal cancellation. The GSS algorithm would attempt to null out those reflections from the unwanted signal. But in addition to nulling out reflections from the unwanted signal, the MMI beamforming algorithm would attempt to *strengthen* those reflections from the desired source; assuming statistically independent sources, strengthening a reflection from the desired source would have little or no effect on the numerator of $|\rho_{12}|^2$, but would increase the denominator, thereby leading to an overall reduction of the optimization criterion. 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 MMI 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 MMI beamforming algorithm forms sidelobes directed towards the reflections of a desired signal, we conducted experiments with a simulated acoustic environment. As shown in Fig. 2, we considered a simple configuration where there are two sound sources, a reflective surface, and an eight-channel linear microphone array that captures both the direct and reflected waves from each source. Actual speech data were used as sound sources in this simulation, which was based on the *image method* [13].

Fig. 3 shows beam patterns at $f_\mathrm{s} = 1500$ Hz and $f_\mathrm{s} = 3000$ Hz obtained with the MMI beamformer and the GSS algorithm. In order to make the techniques directly comparable, the implementation of the GSS algorithm used for the simulation, as well as the ASR experiments described in Section 5, was based on two GSCs, each aimed at one target. Both MMI beamformer and GSS algorithm formed the beam patterns so that the signal from Source 2 in Fig. 2 was enhanced while the other from Source 1 was suppressed. It is clear that both algorithms have unity gain in the look direction, and place deep nulls on the direct path of the unwanted source. The suppression of Reflection 1, the undesired interference, by the MMI beamformer is equivalent to or better than that provided by the

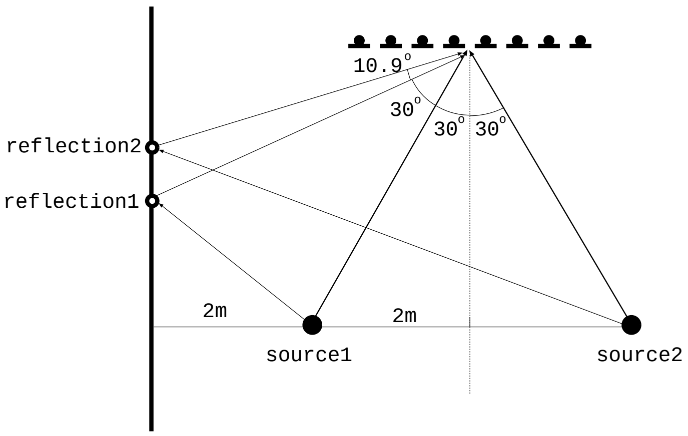
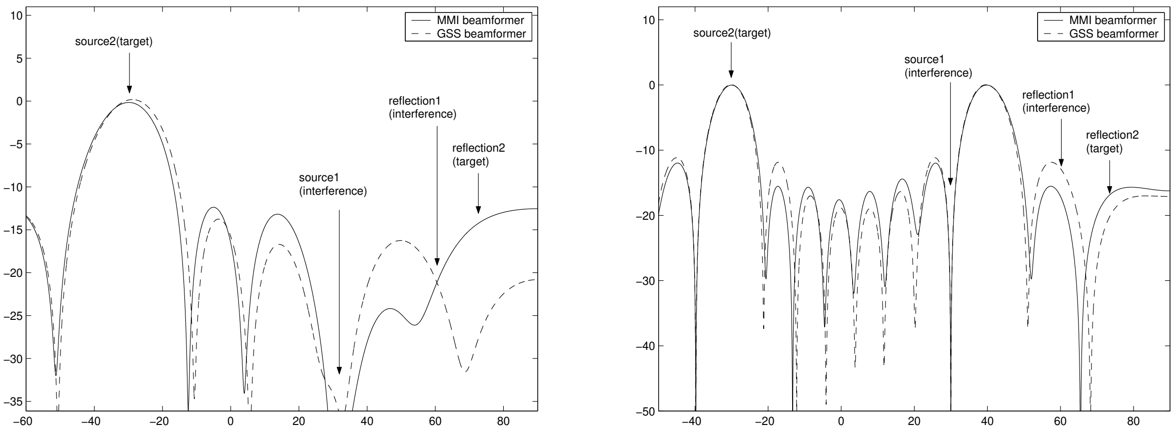


Figure 2: Configuration of sources, sensors, and reflective surface for simulation comparing GSS and MMI beamformer.



(a) (b)

Figure 3: Beam patterns produced by the MMI beamformer and GSS algorithm using a spherical wave assumption for (a) $f_{\rm s} = 1500$ Hz and (b) $f_{\rm s} = 3000$ Hz.

GSS algorithm for both frequencies. Moreover, the enhancement of Reflection 2, the desired signal, by the MMI beamformer is stronger than that of the GSS algorithm.

Given that a beam pattern shows the sensitivity of an array to plane waves, but the beam patterns in Fig. 3 were made with near-field sources and reflections, we also ran a second set of simulations in which all sources and reflections were assumed to produce plane waves. The results of this second simulation are shown in Fig. 4. Once more, it is apparent that the MMI beamformer emphasizes Reflection 2 from the desired source.

If a regularization term is added as before, we obtain the GSS optimization criteria

$$\mathcal{I}'(Y_1, Y_2; \alpha) = |\epsilon_{12}|^2 + \alpha \|\mathbf{w}_{\rm a,1}\|^2 + \alpha \|\mathbf{w}_{\rm a,2}\|^2. \tag{16}$$

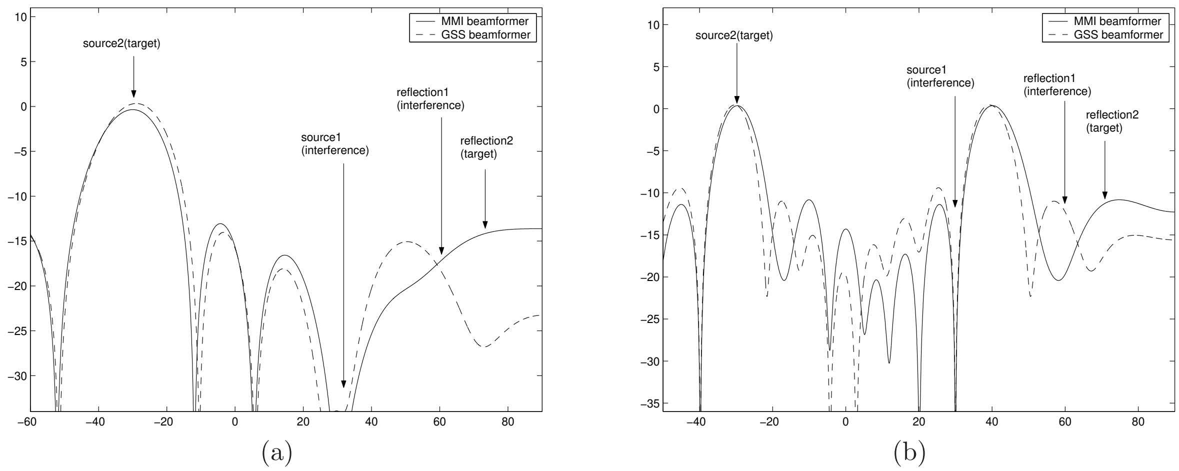


Figure 4: Beam patterns produced by the MMI beamformer and GSS algorithm using a plane wave assumption for (a) $f_{\mathrm{s}} = 1500$ Hz and (b) $f_{\mathrm{s}} = 3000$ Hz.

Then taking partial derivatives of (16) gives

$$\frac{\mathcal{I}'(Y_1, Y_2; \alpha)}{\partial \mathbf{w}_{\mathrm{a},1}^*} = -\mathbf{B}_1^H \Sigma_{\mathbf{X}} (\mathbf{w}_{\mathrm{q},2} - \mathbf{B}_2 \mathbf{w}_{\mathrm{a},2})\, \epsilon_{12}^* + \alpha \mathbf{w}_{\mathrm{a},1} \tag{17}$$

$$\frac{\mathcal{I}'(Y_1, Y_2; \alpha)}{\partial \mathbf{w}_{\mathrm{a},2}^*} = -\mathbf{B}_2^H \Sigma_{\mathbf{X}} (\mathbf{w}_{\mathrm{q},1} - \mathbf{B}_1 \mathbf{w}_{\mathrm{a},1})\, \epsilon_{12} + \alpha \mathbf{w}_{\mathrm{a},2}. \tag{18}$$

Although at first blush it may seem that a closed-form solution for $\mathbf{w}_{\mathrm{a},1}$ and $\mathbf{w}_{\mathrm{a},2}$ could be derived, the presence of ϵ_{12}^* and ϵ_{12} in (17) and (18) respectively actually makes this impossible. Hence, a numerical optimization algorithm is needed, as before.

4 Calculation of Mutual Information 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. 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 mutual information between the outputs of two GSCs.

In the field of *independent component analysis* (ICA), it is common practice to use mutual information as a measure of the independence of two or more signals, as in the prior sections. The entire field of ICA, however, 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 [4]. Briefly, their reasoning is grounded on two points:

1. The *central limit theorem* states that the pdf of the sum of independent 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 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. *Entropy* is the basic measure of information in *information theory* [7]. It is well known that a Gaussian r.v. has the highest entropy of all r.v.s with a given variance [7, Thm. 7.4.1], which holds also for complex Gaussian r.v.s [14, Thm. 2]. Hence, a Gaussian r.v. is, in some sense, the

Table 1: Average log-likelihoods of subband speech samples for various pdfs.

pdf	$\frac{1}{TK}\sum_{t=0}^{T-1}\sum_{m=0}^{M-1}\log p(X_{t,m};\mathrm{pdf})$
Γ	–0.779
K_0	–1.11
Laplace	–2.48
Gaussian	–9.93

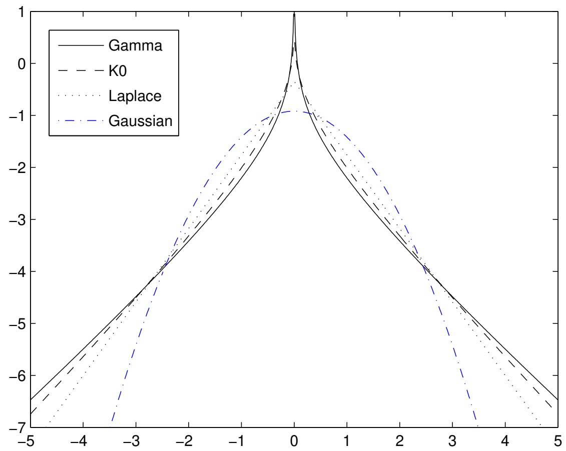


Figure 5: Plot of the log-likelihood of the super-Gaussian and Gaussian pdfs.

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.

Table 1 shows the average log-likelihood of subband samples of speech recorded with a close-talking microphone as calculated with the Gaussian and three super-Gaussian pdfs, namely, the Laplace, K_0, and Γ pdfs averaged over $T = 1000$ time instants and $M = 512$ subbands. It is clear from these log-likelihood values that the complex subband samples of speech are in fact better modeled by the super-Gaussian pdfs considered here than the Gaussian. Hence, the abstract arguments on which the field of ICA are founded correspond well to the actual characteristics of speech. It is worth noting that the use of *spherically-invariant random processes* (SIRPs) in the context of BSS is discussed by Buchner *et al.* [3].

A plot of the log-likelihood of the Gaussian and three super-Gaussian *real* univariate pdfs considered here is provided in Fig. 5. From the figure, it is clear that the Laplace, K_0, and Γ 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.

The *kurtosis* of a r.v. Y, defined as

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

is a measure of how *non-Gaussian* it is [4]. 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 considered here, the Γ pdf has the highest kurtosis, followed by the K_0, then by the Laplace pdf. This fact manifests itself in Fig. 5, where it is clear that as the kurtosis increases, the pdf becomes more and more spikey and heavy-tailed. It is also clear from Table 1 that the average log-likelihood of the subband samples of speech increases significantly as the kurtosis of the pdf used to measure the log-likelihood increases. This is a further proof of the validity of the assumptions on which ICA is based for speech processing.

As explained in Brehm and Stammler [15], it is useful to assume that the Laplace, K_0, and Γ pdfs belong to the class of 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-functions* 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.

For complex Laplace r.v.s $Y_i \in \mathbf{C}$, the univariate pdf can be expressed as

$$p_{\text{Lap}}(Y_i) = \frac{4}{\sqrt{\pi}\sigma_Y^2}\, K_0\left(\frac{2\sqrt{2}|Y_i|}{\sigma_Y}\right) \tag{19}$$

where $K_0(z)$ is an irregular modified Bessel function and $\sigma_Y^2 = \mathcal{E}\{|Y_i|^2\}$. For $\mathbf{Y} \in \mathbf{C}^2$, the bivariate Laplace pdf is given by

$$p_{\text{Lap}}(\mathbf{Y}) = \frac{16}{\pi^{3/2}|\Sigma_{\mathbf{Y}}|\sqrt{s}}\, K_1\left(4\sqrt{s}\right) \tag{20}$$

where $\Sigma_{\mathbf{Y}} = \mathcal{E}\{\mathbf{Y}\mathbf{Y}^H\}$ and

$$s = \mathbf{Y}^H \Sigma_{\mathbf{Y}}^{-1} \mathbf{Y}.$$

Similarly, we can write the univarite K_0 pdf for complex r.v.s $Y_i \in \mathbf{C}$ as

$$p_{K_0}(Y_i) = \frac{1}{\sqrt{\pi}\sigma_Y|Y_i|}\exp\left(-2\,|Y_i|/\sigma_Y\right). \tag{21}$$

The bivariate K_0 pdf for $\mathbf{Y} \in \mathbf{C}^2$ can be expressed as

$$p_{K_0}(\mathbf{Y}) = \frac{\sqrt{2} + 4\sqrt{s}}{2\pi^{3/2}\,|\Sigma_{\mathbf{Y}}|\, s^{3/2}}\exp\left(-2\sqrt{2\,s}\right). \tag{22}$$

Derivations of (19)–(22) are provided in the Appendix. For the Γ pdf, the complex univariate and bivariate pdfs *cannot* be expressed in closed form in terms of elementary or even special functions. As explained in the Appendix, however, it is possible to derive Taylor series expansions that enable the required variates to be calculated to arbitrary accuracy.

The mutual information can no longer be expressed in closed form as in (5) for the super-Gaussian pdfs. We can, however, replace the exact mutual information with the *empirical mutual information*

$$\begin{aligned} I(Y_1, Y_2) \approx \frac{1}{N}\sum_{t=0}^{N-1} \Bigg[&\log p(\mathbf{Y}^{(t)}) \\ &- \sum_{i=1}^{2} \log p(Y_i^{(t)}) \Bigg]. \end{aligned} \tag{23}$$

The relations necessary to evaluate the partial derivative of (23) with respect to $\mathbf{w}_{\text{a},i}$ for the super-Gaussian pdfs considered here are given in [16, §4.6]. Such an empirical approximation was used for the experiments described in the next section.

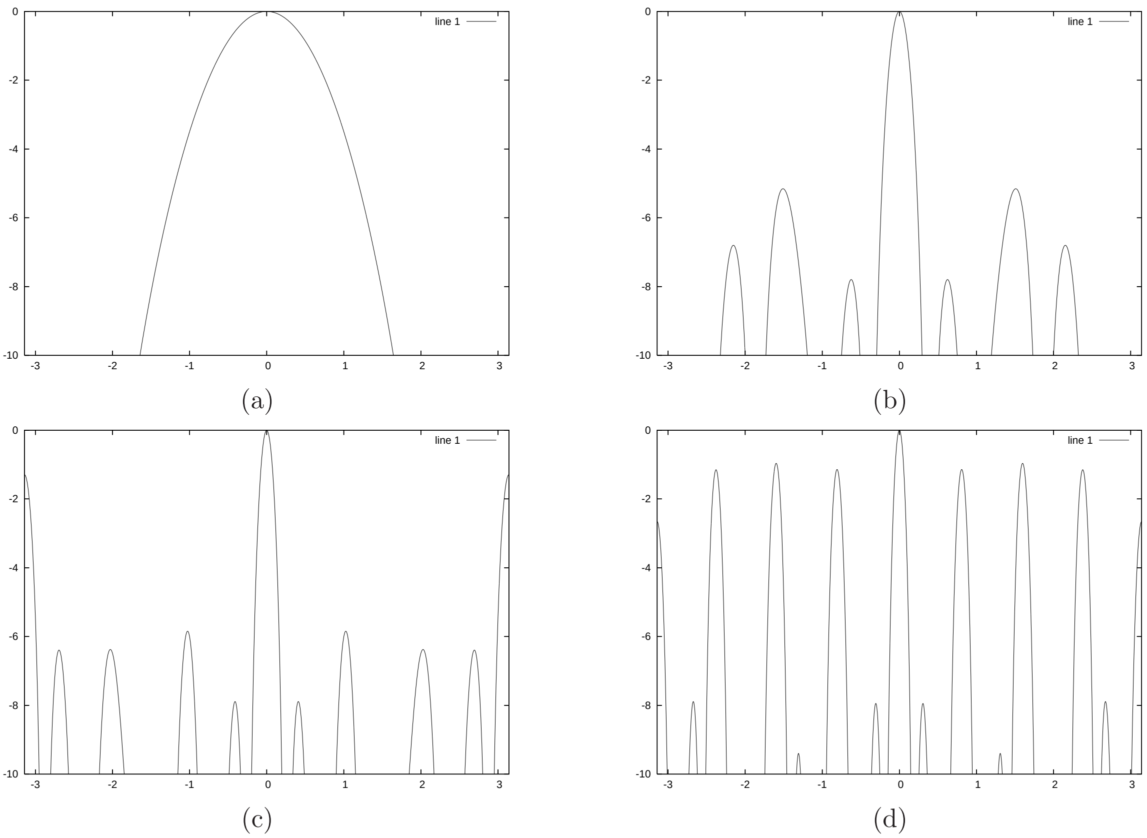


Figure 6: Delay-and-sum beam patterns: Suppression (dB) vs. Azimuth (grad) for (a) λ_s = 50 cm, f_s = 686 Hz; (b) λ_s = 10 cm, f_s = 3,430 Hz; (c) λ_s = 6.67 cm, f_s = 5,145 Hz; (d) λ_s = 5 cm, f_s = 6,860 Hz.

5 Experiments

We performed far-field automatic speech recognition (ASR) experiments on development data from the *PASCAL Speech Separation Challenge* (SSC); see Lincoln *et al.* [5] for a description of the data collection apparatus. The data set contains recordings of five pairs of speakers where each pair of speakers reads approximately 30 sentences taken from the 5,000 word vocabulary Wall Street Journal (WSJ) task. There are a total of 43.9 minutes of speech in the development set and a total of 11,598 word tokens in the reference transcriptions. The data from two simultaneously active speakers was recorded with two circular, eight-channel microphone arrays. The diameter of each array was 20 cm, and the sampling rate of the recordings was 16 kHz. This is a challenging task for source separation algorithms given that the room is reverberant and some recordings include significant amounts of background noise. In addition, as the recorded data is real and not artificially convolved with measured room impulse responses, the position of the speaker's head as well as the speaking volume varies.

Beam patterns obtained with the circular array and delay-and-sum beamforming are shown in Fig. 6 for various *source* wavelengths λ_s and frequencies f_s. These plots were obtained by assuming that a plane wave is propagating parallel to the plane of the circular array. As is clear from Fig. 6 (a), 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. From Fig. 6 (c) and (d), it is clear that 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* [1, §2.5].

Prior to beamforming, we first estimated the speaker's position with the *Orion* source tracking system [17], which is developed and maintained entirely by the second author. In addition to the speaker's position, Orion is also capable of determining when each speaker is active. This information proved very useful segmenting the utterances of each speaker, given that an utterance spoken by one speaker was often much longer than that spoken by the other. In the absence of perfect separation, which we could *not* achieve with the algorithms described here, running the speech recognizer over the entire waveform produced by the beamformer instead of only that portion where a given speaker was actually active would have resulted in significant insertion errors. These insertions would also have proven disastrous for speaker adaptation, as the adaptation data from one speaker would have been contaminated with speech of the other speaker.

Based on the average speaker position estimated for each utterance, utterance-dependent active weight vectors $\mathbf{w}_{a,i}$ were estimated for each source $i = 1, 2$. The active weights for each subband were initialized to zero for estimation with the Gaussian pdf. The snapshot covariance matrix $\Sigma_{\mathbf{X}}$ was estimated for an entire utterance. This matrix was all that was required to estimate $\{\mathbf{w}_{\mathrm{a},i}\}$ for the Gaussian case. For estimation with the super-Gaussian pdfs, the active weights were initialized to their optimal values under the Gaussian assumption. Thereafter iterations of the conjugate gradients algorithm were run on the entire utterance until convergence was achieved.

After beamforming, the feature extraction of our ASR system was based on cepstral features estimated with a warped *minimum variance distortionless response* [18] (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 STC transform [19].

The far-field ASR experiments reported here were conducted with the *Millenium* automatic speech recognition system, which is developed and maintained entirely at the University of Karlsruhe by the fourth and fifth authors. Millenium is based on the *Enigma* weighted finite-state transducer (WFST) library, which contains implementations of all standard WFST algorithms, including weighted composition, weighted determinization, weight pushing, and minimization [20]. The *word trace decoder* in Millenium is implemented along the lines suggested by Saon *et al.* [21], and is capable of generating word lattices, which can then be optimized with WFST operations as in [22]; 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. In addition to the word trace decoder, Millenium also contains a *state trace decoder*, which maintains the full alignment of acoustic features to states during decoding and lattice generation. This state trace decoder is useful for both speaker adaptation and *hidden Markov model* (HMM) parameter estimation.

The training data used for the experiments reported here was taken from the ICSI, NIST, and CMU *meeting corpora*, as well as the *Transenglish Database* (TED) corpus, for a total of 100 hours of training material. In addition to these corpora, approximately 12 hours of speech from the WSJ-CAM0 corpus [23] was used for HMM training in order to provide coverage of the British accents for the speakers in the SSC development set [5]. Acoustic models estimated with two different HMM training schemes were used for several decoding passes: conventional maximum likelihood (ML) HMM training [24, §12], and speaker-adapted training under a ML criterion (ML-SAT) [25]. Our baseline system was fully continuous with 3,500 codebooks and a total of 180,656 Gaussian components.

We performed the 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, language

Table 2: Word error rates for every beamforming algorithm after every decoding passes, as well as the close-talking microphone (CTM).

Beamforming Algorithm	Pass (%WER)			
	1	2	3	4
Delay & Sum	85.1	77.6	72.5	70.4
GSS	80.1	65.5	60.1	56.3
MMI: Gaussian	79.7	65.6	57.9	55.2
MMI: Laplace	81.1	67.9	59.3	53.8
MMI: K_0	78.0	62.6	54.1	52.0
MMI: Γ	80.3	63.0	56.2	53.8
CTM	37.1	24.8	23.0	21.6

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 [26]. A description of the four decoding passes follows:

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

2. Estimate vocal tract length normalization (VTLN) [27] parameters and constrained maximum likelihood linear regression parameters (CMLLR) [28] for each speaker, then redecode with the conventional ML acoustic model and bigram LM.

3. Estimate VTLN, CMLLR, and maximum likelihood linear regression (MLLR) [29] 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 2 shows the word error rate (WER) for each beamforming algorithm after every decoding pass on the SSC data. After the fourth pass, the delay-and-sum beamformer has the worst recognition performance of 70.4% WER. This is not surprising given that the mixed speech was not well separated by the delay-and-sum beamformer for the reasons mentioned above. The WER achieved by the MMI beamformer with a Gaussian pdf of 55.2% was somewhat better than the 56.3% WER from GSS algorithm, which is what should be expected given the reasoning in Section 3.3. The best performance of 52.0% WER was achieved with the MMI beamformer by assuming the subband samples are distributed according to the K_0 pdf.

The WER of 52.0% achieved with the best beamforming algorithm is still more than double the WER of 21.6% achieved with the close-talking microphone (CTM). Hence, there is still a great need for further research to reduce the WER obtained with the separated speech to that obtained with the CTM. A WER of 15–20% is sufficient for a variety of applications including audio indexing; a WER of over 50%, on the other hand, would lead to greatly degraded performance.

Although the Γ pdf assumption gave the highest log-likelihood, as reported in Table 1, the K_0 pdf achieved the best recognition performance. There are several possible explanations for this: Firstly, as mentioned in Section 6, the subband filter bank used for the experiments reported here may not be optimally suited for beamforming and adaptive filtering applications [11]. Hence, aliasing introduced by the filter bank could be masking the gain that would otherwise be obtained by using a pdf with higher kurtosis to calculate mutual information and optimize the active weight vectors. Secondly, data recorded in real environments contains background noise as well as speech. If the pdf of the noise is super-Gaussian, it could conceivably be emphasized by the MMI beamformer with a super-Gaussian pdf assumption. Feature and model adaptation algorithms such as CMLLR and MLLR can, however, robustly estimate parameters to compensate for the background noise. As a result, such an effect is

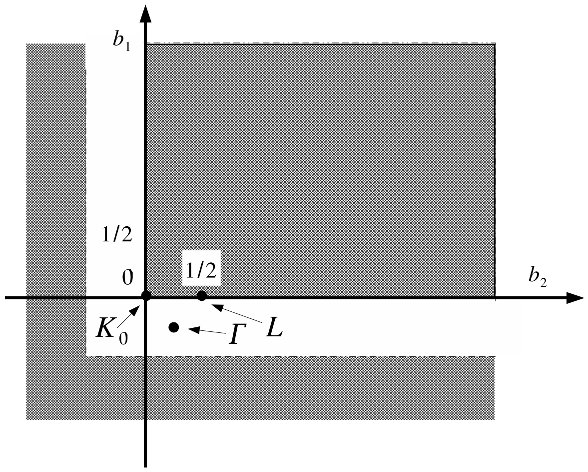


Figure 7: The b_1-b_2 parameter space of $G_{0\,2}^{2\,0}(\lambda x^2|b_1, b_2)$. The unshaded region yields valid pdfs; after Brehm and Stammler [15].

mitigated by the speaker adaptation. From Table 2, this is evident from the significant improvement after the second pass when the Γ pdf is used; to wit, the results obtained with the Γ pdf go from being somewhat worse than the Gaussian results after the first unadapted pass to significantly better after the second pass with VTLN and CMLLR adaptation, and remain significantly better after all subsequent adapted passes.

6 Conclusions and Future Work

In this work, we have proposed a novel beamforming algorithm for simultaneous active speakers based on minimizing mutual information. The proposed method does not exhibit the signal cancellation problems typically seen in conventional adaptive beamformers. Moreover, unlike conventional BSS techniques, the proposed algorithm does not have permutation and scaling ambiguities that cause distortion in the output speech. We evaluated the Gaussian and three super-Gaussian pdfs in calculating the mutual information of the beamformer outputs, and found that the K_0 pdf provides the best ASR performance on the separated speech.

As discussed in the Appendix, the form of the Meijer G-function that is useful for modeling the statistics of speech can be expressed as

$$p_1(y) = A\, G_{0\,2}^{2\,0}(\lambda x^2|b_1, b_2)$$

where the normalization constants A and λ and determined by the real parameters b_1 and b_2. Fig. 7 shows the region of the b_1-b_2 parameter space that yields valid pdfs, along with the points corresponding to the Laplace, K_0, and Γ pdfs [15]. In this work, we have explored only three points in the entire b_1-b_2 space, largely because these points have *names*. One might easily imagine that other pdfs *without names* exist that model the statistics of speech better than those investigated here. In future work, we will systematically search the b_1-b_2 space to find the optimal pdf using a ML criterion.

De Haan *et al* [11] observe that a DFT filter bank based on a single prototype impulse response designed to satisfy a *paraunitary constraint* [10, §8] and thereby achieve perfect reconstruction, such as that used for the experiments reported in Section 5, may not be optimally suited for applications involving beamforming and adaptive filtering. This follows from the fact that the PR design is based on the concept of *aliasing cancellation* [10, §5], whereby the aliasing that is perforce present in a given subband is cancelled out by the aliasing in all other subbands. Aliasing cancellation only works, however, if arbitrary magnitude scale factors and phase shifts are *not* applied to the individual subbands, which is exactly what happens in beamforming and adaptive filtering. The solution proposed by de Haan *et al* [11] is to give up on achieving perfect reconstruction, rather design an analysis prototype so as to minimize the *inband aliasing*, then to design a separate synthesis prototype to minimize a weighted combination of the total response and aliasing distortion. Moreover, they demonstrate that both distortions can be greatly reduced through *oversampling*. In future, we plan to investiage such oversampled DFT filter bank designs.

We also plan to develop an on–line version of the MMI 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.

A Super-Gaussian Distributions

As explained in Brehm and Stammler [15], it is useful to assume that the Laplace, K_0, and Γ pdfs belong to the class of *spherically invariant random processes* (SIRPs) for two principal reasons. Firstly, this implies that multivariates of all orders can be derived from the univariate pdf as soon as the covariance matrix is known; this is most readily accomplished using the formalism of the Meijer G-function. Secondly, such variants can be extended to the case of complex r.v.s, which is essential for our current development. In this appendix, we provide a brief exposition of the Meijer G-function and its use in deriving multivariate super-Gaussian pdfs for complex r.v.s.

A.1 Meijer G-functions

In this section, we very briefly introduce the notation of the Meijer G-function, along with the most important relations required to use G-functions to model super-Gaussian pdfs.

To denote the Meijer G-function, we will use one of the following equivalent forms

$$
\begin{aligned}
&G^{m\,n}_{p\,q}\left(z \,\middle|\, \begin{matrix} a_p \\ b_q \end{matrix}\right) \\
&\quad = G^{m\,n}_{p\,q}\left(z \,\middle|\, \begin{matrix} a_1,\ldots,a_p \\ b_1,\ldots,b_q \end{matrix}\right) \\
&\quad = G^{m\,n}_{p\,q}\left(z \,\middle|\, \begin{matrix} a_1,\ldots,a_n & | & a_{n+1},\ldots,a_p \\ b_1,\ldots,b_m & | & b_{m+1},\ldots,b_q \end{matrix}\right).
\end{aligned}
$$

The G-function is defined by the contour integral

$$
\begin{aligned}
G^{m\,n}_{p\,q}\left(x \,\middle|\, \begin{matrix} a_1,\ldots,a_p \\ b_1,\ldots,b_q \end{matrix}\right) &= \frac{1}{2\pi i}\oint_{\Gamma_L} x^s ds \\
&\times \frac{\prod_{j=1}^{m}\Gamma(b_j - s)\prod_{j=1}^{n}\Gamma(1-a_j+s)}{\prod_{j=n+1}^{p}\Gamma(a_j - s)\prod_{j=m+1}^{q}\Gamma(1-b_j+s)}
\end{aligned}
\tag{24}
$$

where $\Gamma(z)$ is the Gamma function and Γ_L is a contour of integration defined as in [15]. The definition (24) implies

$$
G^{m\,n}_{p\,q}\left(z \,\middle|\, \begin{matrix} a_p \\ b_q \end{matrix}\right) = z^{-u} G^{m\,n}_{p\,q}\left(z \,\middle|\, \begin{matrix} a_p + u \\ b_q + u \end{matrix}\right) \tag{25}
$$

where $a_p + u$ and $b_q + u$ indicate that u is to be added to all $a_1, \ldots, a_p$ and all $b_1, \ldots, b_q$, respectively. To determine the normalizing constants of the several pdfs generated from the Meijer G-function, it will be useful to apply the Mellin transform

$$M\{f(x); z\} = \int_0^\infty dx\, x^{z-1}\, f(x). \tag{26}$$

Under suitable conditions [15], the Mellin transform of a Meijer G-function can be expressed as

$$\begin{aligned} M & \left\{ G_{p\,q}^{m\,n} \left(z \middle| \begin{matrix} a_p \\ b_q \end{matrix} \right); z \right\} \\ & = \frac{\prod_{i=1}^{m} \Gamma(b_i + z) \prod_{i=1}^{n} \Gamma(1 - a_i - z)}{\prod_{i=1}^{m} \Gamma(1 - b_i - z) \prod_{i=1}^{n} \Gamma(a_i + z)}. \end{aligned} \tag{27}$$

A.2 Spherically Invariant Random Processes

We now show how G-functions can be used to represent SIRPs. To begin, we can express a univariate pdf of a SIRP as

$$p_1(x) = A G_{p\,q}^{m\,n} \left(\lambda\, x^2 \middle| \begin{matrix} a_p \\ b_q \end{matrix} \right) \tag{28}$$

for all $-\infty < x < \infty$. As can be verified by the Mellin transform relations (26)–(27), the normalization factor A and the constant λ, which assures unity variance, must be chosen according to

$$A = \lambda^{1/2} \frac{\prod_{i=m+1}^{q} \Gamma(\frac{1}{2} - b_i) \prod_{i=n+1}^{p} \Gamma(\frac{1}{2} + a_i)}{\prod_{i=i}^{m} \Gamma(\frac{1}{2} + b_i) \prod_{i=1}^{n} \Gamma(\frac{1}{2} - a_i)} \tag{29}$$

$$\lambda = (-1)^\epsilon \frac{\prod_{i=1}^{q} (\frac{1}{2} + b_i)}{\prod_{i=1}^{p} (\frac{1}{2} + a_i)}, \qquad \epsilon = n - (q - m). \tag{30}$$

Brehm and Stammler [15] note that the subclass of SIRPs that are useful for modeling the statistics of speech can be expressed as

$$p_1(y) = A\, G_{0\,2}^{2\,0}(\lambda x^2 | b_1, b_2) \tag{31}$$

for the real parameters b_1 and b_2, where (29–30) are specialized as

$$A = \frac{\lambda^{1/2}}{\Gamma(\frac{1}{2} + b_1)\Gamma(\frac{1}{2} + b_2)} \tag{32}$$

and

$$\lambda = (\tfrac{1}{2} + b_1)(\tfrac{1}{2} + b_2). \tag{33}$$

Table 3, taken from Brehm and Stammler [15], lists the values of these parameters for the Laplace, K_0, and Γ pdfs. In many cases of interest, a Meijer G-function with a given set of parameters

Table 3: Meijer G-function parameter values for the Laplace, K_0, and Γ pdfs.

pdf	$p(x)$	b_1	b_2	A	λ
Laplace	$\frac{1}{\sqrt{2}}e^{-\sqrt{2}\|x\|}$	0	$\frac{1}{2}$	$(2\pi)^{-1/2}$	$\frac{1}{2}$
K_0	$\frac{1}{\pi}K_0(\|x\|)$	0	0	$(2\pi)^{-1}$	$\frac{1}{4}$
Γ	$\frac{\sqrt{3}}{4\sqrt{\pi}}\left(\frac{\sqrt{3}\|x\|}{2}\right)^{-1/2} e^{-\sqrt{3}\|x\|/2}$	-$\frac{1}{4}$	$\frac{1}{4}$	$\frac{\sqrt{3/2}}{4\pi}$	$\frac{3}{16}$

can be represented in closed-form in terms of elementary or special functions. These special cases are tabulated in reference books such as Luke [6]. Alternatively, they have been programmed into computer algebra systems, such as *Mathematica* [30, §3.2.10]. In particular, we can write

$$G^{2\,0}_{0\,2}(z|0, \tfrac{1}{2}) = \sqrt{\pi}e^{-2\sqrt{z}} \tag{34}$$

$$G^{2\,0}_{0\,2}(z|0, 0) = 2\,K_0(2\sqrt{z}). \tag{35}$$

These equations can be used to verify the correctness of the Laplace and K_0 pdfs. To verify the correctness of the Γ density, we write

$$G^{2\,0}_{0\,2}(z|-\tfrac{1}{4}, \tfrac{1}{4}) = z^{-1/4}G^{2\,0}_{0\,2}(z|0, \tfrac{1}{2}) \tag{36}$$

$$= \sqrt{\pi}\,z^{-1/4}\,e^{-2\sqrt{z}} \tag{37}$$

where (36) follows from (25), and (37) follows from (34).

In general, the multivariate density of order ν can also be expressed in terms of Meijer's G-functions according to [15]

$$p_\nu(\mathbf{x}) = \pi^{-\nu/2} f_\nu(s) \tag{38}$$

where

$$\begin{aligned} f_\nu &= \pi^{1/2}\,A_\nu\,s^{(1-\nu)/2} \\ &\times G^{3\,0}_{1\,3}\left(\lambda_\nu s \,\middle|\, \begin{matrix} 0 \\ \frac{1}{2}(\nu-1), b_1, b_2 \end{matrix}\right) \end{aligned} \tag{39}$$

and $s = \mathbf{x}^T\mathbf{x}$. In this case (29) and (30) can be specialized as

$$\epsilon = 0$$

$$A_\nu = \lambda_\nu^{1/2}\frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{1}{2}\nu)\Gamma(\frac{1}{2}+b_1)\,\Gamma(\frac{1}{2}+b_2)} \tag{40}$$

$$\lambda_\nu = \nu\,(\tfrac{1}{2}+b_1)\,(\tfrac{1}{2}+b_2). \tag{41}$$

The bivariate pdf is obtained by specializing (38) and (39) as,

$$p_2(\mathbf{x}) = \frac{A_2}{\sqrt{\pi s}}\,G^{3\,0}_{1\,3}\left(\lambda_2 s \,\middle|\, \begin{matrix} 0 \\ \frac{1}{2}, b_1, b_2 \end{matrix}\right). \tag{42}$$

For the moment, assume $\mathbf{x}$ is real-valued; this analysis will be extended to the case of complex $\mathbf{x}$ in Section A.6. If the components of $\mathbf{x}$ are correlated, we must set

$$s = \mathbf{x}^T\Sigma_{\mathbf{X}}{}^{-1}\mathbf{x}$$

and modify (42) according to

$$p_2(\mathbf{x}) = \frac{A_2}{\sqrt{\pi s|\Sigma_{\mathbf{X}}|}}\,G^{3\,0}_{1\,3}\left(\lambda_2 s \,\middle|\, \begin{matrix} 0 \\ \frac{1}{2}, b_1, b_2 \end{matrix}\right) \tag{43}$$

where $\Sigma_{\mathbf{X}} = \mathcal{E}\{\mathbf{X}\mathbf{X}^T\}$ is the covariance matrix of $\mathbf{X}$.

For the four-variate case, we have

$$p_4(\mathbf{x}) = \frac{A_4}{(\pi s)^{3/2}\,|\Sigma_{\mathbf{X}}|^{1/2}}\, G^{3\,0}_{1\,3}\left(\lambda_4 s \,\middle|\, \begin{matrix} 0 \\ \frac{3}{2}, b_1, b_2 \end{matrix}\right). \tag{44}$$

A.3 Laplace Density

The Laplace density is perhaps the simplest and best known super-Gaussian distribution. In Table 3, the univariate form of the Laplace density is given, along with the parameter values required to represent it with Meijer's G-function as in (31–33). With the help of Mathematica, we learn

$$G^{3\,0}_{1\,3}\left(z \,\middle|\, \begin{matrix} 0 \\ \frac{1}{2}, 0, \frac{1}{2} \end{matrix}\right) = 2\sqrt{z}\,K_0(2\sqrt{z}).$$

Hence, specializing (43) with $b_1 = 0$ and $b_2 = \frac{1}{2}$, then simplifying provides the bivariate pdf

$$p_2(\mathbf{x}) = \frac{2A_2\sqrt{\lambda_2}}{\sqrt{\pi|\Sigma_{\mathbf{X}}|}}\, K_0\left(2\sqrt{\lambda_2 s}\right) \tag{45}$$

where from (40–41) we have

$$\lambda_2 = 2\left(\tfrac{1}{2}+0\right)\left(\tfrac{1}{2}+\tfrac{1}{2}\right) = 1 \tag{46}$$

$$A_2 = \frac{\Gamma(\frac{1}{2})}{\Gamma(1)\,\Gamma(\frac{1}{2})\,\Gamma(1)} = \frac{1}{\Gamma^2(1)} = 1. \tag{47}$$

Substituting (46–47) into (45), we have

$$p_2(\mathbf{x}) = \frac{2}{\sqrt{\pi|\Sigma_{\mathbf{X}}|}}\, K_0(2\sqrt{s}). \tag{48}$$

Once more resorting to Mathematica, we find

$$G^{3\,0}_{1\,3}\left(z \,\middle|\, \begin{matrix} 0 \\ \frac{3}{2}, 0, \frac{1}{2} \end{matrix}\right) = 2z\,K_1(2\sqrt{z}).$$

Hence, specializing (44) provides the four-variate pdf

$$p_4(\mathbf{x}) = \frac{2A_4\lambda_4}{\pi^{3/2}\,s^{1/2}\,|\Sigma_{\mathbf{X}}|^{1/2}} K_1\left(2\sqrt{\lambda_4 s}\right) \tag{49}$$

where

$$\lambda_4 = 2\cdot 1 = 2 \tag{50}$$

$$A_4 = \sqrt{2}\cdot\frac{\Gamma(\frac{1}{2})}{\Gamma(2)\,\Gamma(\frac{1}{2})\,\Gamma(1)} = \frac{\sqrt{2}}{\Gamma(2)\Gamma(1)} = \sqrt{2}. \tag{51}$$

Substituting (50–51) back into (49) provides

$$p_4(\mathbf{x}) = \frac{4\sqrt{2}}{\pi^{3/2}\,s^{1/2}\,|\Sigma_{\mathbf{X}}|^{1/2}} K_1(2\sqrt{2s}). \tag{52}$$

A.4 K_0 Density

From Mathematica

$$G^{3\,0}_{1\,3}\left(z\,\middle|\,\begin{matrix}0\\ \frac{1}{2},0,0\end{matrix}\right)=\sqrt{\pi}e^{-2\sqrt{z}}$$

so that the bivariate K_0 pdf can be obtained by substituting $b_1 = b_2 = 0$ into (43), whereupon we find

$$p_2(\mathbf{x}) = \frac{A_2}{\sqrt{s|\Sigma_{\mathbf{X}}|}}\, e^{-2\sqrt{\lambda_2 s}} \tag{53}$$

where

$$\lambda_2 = \tfrac{1}{2} \tag{54}$$

$$A_2 = \frac{\sqrt{2}}{2}\cdot\frac{\Gamma(\frac{1}{2})}{\Gamma(1)\,\Gamma(\frac{1}{2})\,\Gamma(\frac{1}{2})} = \frac{\sqrt{2}}{2\,\Gamma(1)\,\Gamma(\frac{1}{2})} = \frac{1}{\sqrt{2\pi}}. \tag{55}$$

Substituting (54–55) into (53), we find

$$p_2(\mathbf{x}) = \frac{1}{\sqrt{2\pi\, s|\Sigma_{\mathbf{X}}|}}\, e^{-\sqrt{2s}}. \tag{56}$$

From Mathematica

$$G^{3\,0}_{1\,3}\left(z\,\middle|\,\begin{matrix}0\\ \frac{3}{2},0,0\end{matrix}\right)=\frac{\sqrt{\pi}(1+2\sqrt{z})}{2}e^{-2\sqrt{z}},$$

so the four-variate K_0 pdf can be obtained from (44),

$$p_4(\mathbf{x}) = \frac{A_4(1+2\sqrt{\lambda_4 s})}{2\pi s^{3/2}|\Sigma_{\mathbf{X}}|^{1/2}}\, e^{-2\sqrt{\lambda_4\, s}} \tag{57}$$

where

$$\lambda_4 = 2\cdot\tfrac{1}{2} = 1 \tag{58}$$

$$A_4 = \frac{\Gamma(\frac{1}{2})}{\Gamma(2)\,\Gamma(\frac{1}{2})\,\Gamma(\frac{1}{2})} = \frac{1}{\Gamma(2)\,\Gamma(\frac{1}{2})} = \frac{1}{\sqrt{\pi}}. \tag{59}$$

Substituting (58–59) into (57), we have

$$p_4(\mathbf{x}) = \frac{(1+2\sqrt{s})}{2(\pi s)^{3/2}|\Sigma_{\mathbf{X}}|^{1/2}}\, e^{-2\sqrt{s}}. \tag{60}$$

A.5 Γ Density

For the Γ pdf, $b_1 = -\frac{1}{4}$, $b_2 = \frac{1}{4}$. Subsituting these values into the G-functions appearing in (43–44) and applying (25), we find

$$\begin{aligned}&G^{3\,0}_{1\,3}\left(z\,\middle|\,\begin{matrix}0\\ \frac{1}{2},-\frac{1}{4},\frac{1}{4}\end{matrix}\right)=\\ &z^{-1/4}G^{3\,0}_{1\,3}\left(z\,\middle|\,\begin{matrix}\frac{1}{4}\\ \frac{3}{4},0,\frac{1}{2}\end{matrix}\right)\end{aligned} \tag{61}$$

and

$$\begin{aligned}&G^{3\,0}_{1\,3}\left(z\,\middle|\,\begin{matrix}0\\ \frac{3}{2},-\frac{1}{4},\frac{1}{4}\end{matrix}\right)=\\ &z^{-1/4}G^{3\,0}_{1\,3}\left(z\,\middle|\,\begin{matrix}\frac{1}{4}\\ \frac{7}{4},0,\frac{1}{2}\end{matrix}\right).\end{aligned} \tag{62}$$

Then, the bi-variate Γ pdf can be expressed as

$$p_2(\mathbf{x}) = \frac{A_2}{\sqrt{\pi s |\Sigma_{\mathbf{x}}|}} (\lambda_2 s)^{-1/4} g_2(\lambda_2 s) \tag{63}$$

where

$$g_2(z) = G^{3\,0}_{1\,3}\left(z \,\middle|\, \begin{matrix} \frac{1}{4} \\ \frac{3}{4}, 0, \frac{1}{2} \end{matrix} \right) \tag{64}$$

$$\lambda_2 = \frac{(\frac{1}{2}+\frac{1}{2})(\frac{1}{2}-\frac{1}{4})(\frac{1}{2}+\frac{1}{4})}{(\frac{1}{2}+0)} = \frac{3}{8} \tag{65}$$

$$A_2 = \sqrt{\frac{3}{8}} \cdot \frac{\Gamma(\frac{1}{2}+0)}{\Gamma(\frac{1}{2}+\frac{1}{2})\,\Gamma(\frac{1}{2}-\frac{1}{4})\,\Gamma(\frac{1}{2}+\frac{1}{4})} \approxeq 0.2443. \tag{66}$$

For the four-variate Γ pdf, we can write

$$p_4(\mathbf{x}) = \frac{A_4}{(\pi s)^{3/2} \sqrt{|\Sigma_{\mathbf{x}}|}} (\lambda_4 s)^{-1/4} g_4(\lambda_4 s) \tag{67}$$

where

$$g_4(z) = G^{3\,0}_{1\,3}\left(z \,\middle|\, \begin{matrix} \frac{1}{4} \\ \frac{7}{4}, 0, \frac{1}{2} \end{matrix} \right) \tag{68}$$

$$\lambda_4 = \frac{(\frac{1}{2}+\frac{3}{2})(\frac{1}{2}-\frac{1}{4})(\frac{1}{2}+\frac{1}{4})}{(\frac{1}{2}+0)} = \frac{3}{4} \tag{69}$$

$$A_4 = \frac{\sqrt{3}}{2} \cdot \frac{\Gamma(\frac{1}{2}+0)}{\Gamma(\frac{1}{2}+\frac{3}{2})\,\Gamma(\frac{1}{2}-\frac{1}{4})\,\Gamma(\frac{1}{2}+\frac{1}{4})} \approxeq 0.1949. \tag{70}$$

Unfortunately, the G-functions appearing on the R.H.S. of (61–62) cannot be expressed in closed-form in terms of elementary or special functions. Hence, it is necessary to use a series expansion to calculate them. The Taylor series [31, §19] of any function $f(z)$ about $z = z_0$ can be expressed as

$$f(z) = \sum_{n=0}^{\infty} \frac{(z-z_0)^n}{n!} f^{(n)}(z_0)$$

where $f^{(n)}(z)$ indicates the nth derivative of $f(z)$ evaluated at $z = z_0$. For series expansions of G-functions, the relation [6, §5.4]

$$z^k \frac{d^k}{dz^k} \left\{ G^{m\,n}_{p\,q}\left(z^{-1} \,\middle|\, \begin{matrix} a_p \\ b_q \end{matrix} \right) \right\} = (-)^k G^{m\,n+1}_{p+q\,q+1}\left(z^{-1} \,\middle|\, \begin{matrix} 1-k, a_p \\ b_q, 1 \end{matrix} \right)$$

can be used to evaluate the required derivatives. Note that it is not possible to expand the G-function about the origin $z = 0$, as the G-function has a *branch point singularity* at the origin [32, §10.2]. The G-function can, however, be expanded about any point on the positive real axis, which is sufficient for our purposes here.

In practice, a log-likelihood of the Γ pdf is required. Accordingly we need to calculate the logarithm of the G-function. In order to calculate it precisely, the series expansion is performed about 74 points, and we use the series expanded about the point closest to the given argument up to the 12th order. In the case of $s \geqq 70$ in (63) or (67), we use the derivative to the first order, that is, we used a linear approximation in the log domain. This is because the G-function for those values effectively vanishes leading to floating point errors. Table 4 shows the series coefficients when $\log g_2(z)$ and $\log g_4(z)$ are expanded about $z_0 = 1$.

Table 4: Series coefficients of $\log g_2(z)$ and $\log g_4(z)$.

n	$(\log g_2)^{(n)}\,(z=1)$	$(\log g_4)^{(n)}\,(z=1)$
0	0.254766	0.389422
1	-0.198347	-0.17901
2	0.228596	0.0967777
3	-0.382523	-0.0266552
4	0.887333	-0.179479
5	-2.70435	1.17531
6	10.3182	-6.79936
7	-47.3711	42.6283
8	253.441	-299.361
9	-1538.09	2358.89
10	10330.3	-20730.1
11	-74825.6	201601.8
12	565360.5	-2.15304 $\times 10^6$

A.6 Complex Densities

The multivariate pdfs derived thus far have been for real-valued random vectors. In order to extend this development for complex-valued subband samples, we will adapt a theorem proven by Neeser and Massey [14, Appendix].

The following definition is due to Neeser and Massey [14, Appendix].

Definition 1 *The random vector* $\mathbf{Y} \in \mathbf{C}^N$ *is a* proper random vector *if*

$$\mathcal{E}\{\mathbf{Y}\mathbf{Y}^T\} = \mathbf{0}. \tag{71}$$

Neeser and Massey [14] call the matrix on the L.H.S. of (71), the *pseudo-covariance matrix*. Hence, a proper complex random vector is one for which the psuedo-covariance matrix vanishes.

Lemma 1 *Let* $\mathbf{C}^N \ni \mathbf{Y} = \mathbf{X}_c + i\mathbf{X}_s$ *be a proper random vector with pseudo-covariance matrix.*

$$\Sigma_{\mathbf{Y}} = \mathcal{E}\{\mathbf{Y}\mathbf{Y}^T\} = \mathbf{\Sigma}_{cc} - \mathbf{\Sigma}_{ss} + i(\mathbf{\Sigma}_{sc} + \mathbf{\Sigma}_{sc}^T) \tag{72}$$

where

$$\mathbf{\Sigma}_{cc} = \mathcal{E}\{\mathbf{X}_c{\mathbf{X}_c}^T\} \tag{73}$$
$$\mathbf{\Sigma}_{ss} = \mathcal{E}\{\mathbf{X}_s{\mathbf{X}_s}^T\} \tag{74}$$
$$\mathbf{\Sigma}_{sc} = \mathcal{E}\{\mathbf{X}_s{\mathbf{X}_c}^T\}. \tag{75}$$

Then

$$\mathbf{\Sigma}_{cc} = \mathbf{\Sigma}_{ss} \tag{76}$$
$$\mathbf{\Sigma}_{sc} = -\mathbf{\Sigma}_{sc}^T. \tag{77}$$

Proof: The definition of properness requires that the R.H.S. of (72) vanishes, which implies (76) and (77). □

Note that a matrix satisfying (77) is said to be *skew symmetric*. Hence, the conditions (76) and (77) state that the covariance matrices of the real and imaginary parts of a proper complex random vector must be equal, and the cross-covariance matrices must be skew symmetric.

We now state another intermediate result.

Lemma 2 *Let $\mathbf{M}_{cc}$, $\mathbf{M}_{cc}$, $\mathbf{M}_{sc}$, and $\mathbf{M}_{sc}$, be real $N \times N$ matrices, where $\mathbf{M}_{cc}$ and $\mathbf{M}_{cc}$ are symmetric and $\mathbf{M}_{cs}^T = \mathbf{M}_{sc}$. Define the $N \times N$ Hermitian matrix*

$$\mathbf{M} = \mathbf{M}_c + i\mathbf{M}_s \triangleq \mathbf{M}_{cc} + \mathbf{M}_{cc} + i(\mathbf{M}_{sc} - \mathbf{M}_{sc}{}^T) \tag{78}$$

and the symmetric $2N \times 2N$ matrix

$$\Upsilon \triangleq 2 \begin{bmatrix} \mathbf{M}_{cc} & \mathbf{M}_{cs} \\ \mathbf{M}_{sc} & \mathbf{M}_{cc} \end{bmatrix}. \tag{79}$$

Then the quadratic forms

$$\mathcal{E} \triangleq \mathbf{z}^H \mathbf{M} \mathbf{z} \tag{80}$$

and

$$\mathcal{E}' \triangleq \begin{bmatrix} \mathbf{z}_c^T & \mathbf{z}_s^T \end{bmatrix} \Upsilon \begin{bmatrix} \mathbf{z}_c \\ \mathbf{z}_s \end{bmatrix} \tag{81}$$

are equal for all $\mathbf{z} \triangleq \mathbf{z}_c + i\mathbf{z}_s$, if and only if

$$\mathbf{M}_{cc} = \mathbf{M}_{cc} \text{ and } \mathbf{M}_{sc} = -\mathbf{M}_{sc}^T. \tag{82}$$

Moreover, under conditions (82) $\mathbf{M}$ is positive (semi-)definite if and only if Υ is positive (semi-)definite.

Proof: See Neeser and Massey [14, Appendix]. □

We now state and prove the main result of this section based on [14, Appendix].

Theorem 1 *Consider a proper complex random vector*

$$\mathbf{C}^N \ni \mathbf{Y} = \mathbf{X}_c + i\mathbf{X}_s$$

with the covariance matrix

$$\Sigma_{\mathbf{Y}} = 2(\mathbf{\Sigma}_{cc} + i\mathbf{\Sigma}_{sc}) \tag{83}$$

where $\mathbf{\Sigma}_{cc}$ and $\mathbf{\Sigma}_{sc}$ are defined in (73) and (75), respectively. Define the stacked random vector

$$\mathbf{R}^{2N} \ni \mathbf{X} = \begin{bmatrix} \mathbf{X}_c \\ \mathbf{X}_s \end{bmatrix}$$

with covariance matrix

$$\Sigma_{\mathbf{X}} = \mathcal{E}\{\mathbf{X}\,\mathbf{X}^T\} = \begin{bmatrix} \mathbf{\Sigma}_{cc} & \mathbf{\Sigma}_{cs} \\ \mathbf{\Sigma}_{sc} & \mathbf{\Sigma}_{ss} \end{bmatrix}.$$

Then,

$$\mathbf{x}^T \Sigma_{\mathbf{X}}^{-1} \mathbf{x} = 2\,\mathbf{y}^H \Sigma_{\mathbf{Y}}^{-1} \mathbf{y} \tag{84}$$

for all

$$\mathbf{y} = \mathbf{x}_c + i\mathbf{x}_s \text{ and } \mathbf{x} = \begin{bmatrix} \mathbf{x}_c \\ \mathbf{x}_s \end{bmatrix}.$$

Moreover,

$$\sqrt{|\Sigma_{\mathbf{X}}|} = 2^{-N}\,|\Sigma_{\mathbf{Y}}|. \tag{85}$$

Proof: Based on a well-known result for the inverse of block matrices [33, pg. 656], we can write

$$\mathbf{\Sigma}_{\mathbf{X}}^{-1} = \begin{bmatrix} \Delta^{-1} & {\mathbf{\Sigma}_{\rm cc}}^{-1}\mathbf{\Sigma}_{\rm sc}\Delta^{-1} \\ -\Delta^{-1}\mathbf{\Sigma}_{\rm sc}{\mathbf{\Sigma}_{\rm cc}}^{-1} & \Delta^{-1} \end{bmatrix} \tag{86}$$

where

$$\Delta \triangleq \mathbf{\Sigma}_{\rm cc} + \mathbf{\Sigma}_{\rm sc}{\mathbf{\Sigma}_{\rm cc}}^{-1}\mathbf{\Sigma}_{\rm sc} \tag{87}$$

is symmetric. We must now show that the upper-right block of $\mathbf{\Sigma}_{\mathbf{X}}^{-1}$ is skew symmetric. Observe that

$$\Delta{\mathbf{\Sigma}_{\rm cc}}^{-1}\mathbf{\Sigma}_{\rm sc} = \mathbf{\Sigma}_{\rm sc} + \mathbf{\Sigma}_{\rm sc}{\mathbf{\Sigma}_{\rm cc}}^{-1}\mathbf{\Sigma}_{\rm sc}{\mathbf{\Sigma}_{\rm cc}}^{-1}\mathbf{\Sigma}_{\rm sc} = \mathbf{\Sigma}_{\rm sc}{\mathbf{\Sigma}_{\rm cc}}^{-1}\Delta$$

which implies

$$\begin{aligned} {\mathbf{\Sigma}_{\rm cc}}^{-1}\mathbf{\Sigma}_{\rm sc}\Delta^{-1} &= \Delta^{-1}\mathbf{\Sigma}_{\rm sc}{\mathbf{\Sigma}_{\rm cc}}^{-1} = \left({\mathbf{\Sigma}_{\rm cc}}^{-1}{\mathbf{\Sigma}_{\rm sc}}^{T}\Delta^{-1}\right)^{T} \\ &= -\left({\mathbf{\Sigma}_{\rm cc}}^{-1}\mathbf{\Sigma}_{\rm sc}\Delta^{-1}\right)^{T}. \end{aligned}$$

Hence, the upper and lower blocks are skew symmetric. Therefore, $\mathbf{\Sigma}_{\mathbf{X}}^{-1}$ satisfies (82) and Lemma 2 applies for $\Upsilon \triangleq \frac{1}{2}\mathbf{\Sigma}_{\mathbf{X}}^{-1}$ and

$$\mathbf{M} \triangleq \Delta^{-1}(\mathbf{I} - i\mathbf{\Sigma}_{\rm sc}{\mathbf{\Sigma}_{\rm cc}}^{-1}) \tag{88}$$

where (88) follows from associating the block components in (79) with their counterparts in (86), then applying (78). Multiplying $\mathbf{M}$ in (88) with (83) yields the identity matrix, which implies $\mathbf{M} = \mathbf{\Sigma}_{\mathbf{Y}}^{-1}$. Therefore (84) follows from Lemma 1.

Using a well-known result on the determinant of block matrices [33, pg. 650] and the skew symmetry of $\mathbf{\Sigma}_{cs}$, we find

$$|\Sigma_{\mathbf{X}}| = |\mathbf{\Sigma}_{\rm cc}|\,|\Delta|. \tag{89}$$

Observe that

$$\begin{aligned} \mathbf{\Sigma}_{\mathbf{Y}}^{T} &= 2(\mathbf{\Sigma}_{\rm cc} - i\mathbf{\Sigma}_{\rm sc}) \\ &= 2(\mathbf{I} - i\mathbf{\Sigma}_{\rm sc}{\mathbf{\Sigma}_{\rm cc}}^{-1})\mathbf{\Sigma}_{\rm cc}. \end{aligned} \tag{90}$$

Hence, from (88) and (90), along with $\mathbf{M} = \mathbf{\Sigma}_{\mathbf{Y}}^{-1}$, it follows that

$$\mathbf{\Sigma}_{\mathbf{Y}}^{-1} = \tfrac{1}{4}\Delta^{-1}\mathbf{\Sigma}_{\mathbf{Y}}^{T}{\mathbf{\Sigma}_{\rm cc}}^{-1} \tag{91}$$

Now

$$\left|\Sigma_{\mathbf{Y}}\mathbf{\Sigma}_{\mathbf{Y}}^{-1}\right| = \left|\tfrac{1}{4}\Sigma_{\mathbf{Y}}\Delta^{-1}\mathbf{\Sigma}_{\mathbf{Y}}^{T}{\mathbf{\Sigma}_{\rm cc}}^{-1}\right| \tag{92}$$

$$= \frac{|\Sigma_{\mathbf{Y}}|^2}{2^{2N}\,|\Delta|\,|\mathbf{\Sigma}_{\rm cc}|} = 1 \tag{93}$$

where (92) follows from (91), and (93) follows from a basic property of determinants of matrices. Substituting as in (89) for $|\Delta|\,|\mathbf{\Sigma}_{\rm cc}|$ in (93) and rearranging is sufficient to prove (85). $\square$

Based on Theorem 1, we can rewrite (48) for proper $y \in \mathbf{C}$ as

$$p_{\rm Laplace}(y) = \frac{4}{\sqrt{\pi}\sigma_Y^2}\,K_0\left(\frac{2\sqrt{2}|y|}{\sigma_Y}\right) \tag{94}$$

where $\sigma_Y^2 = \mathcal{E}\{|Y|^2\}$. For proper $\mathbf{y} \in \mathbf{C}^2$, we can rewrite (52) as

$$p_{\rm Laplace}(\mathbf{y}) = \frac{16}{\pi^{3/2}s^{1/2}|\Sigma_{\mathbf{Y}}|}\,K_1\left(4\sqrt{s}\right) \tag{95}$$

where $\Sigma_{\mathbf{Y}} = \mathcal{E}\{\mathbf{Y}\mathbf{Y}^H\}$ and

$$s = \mathbf{y}^H \Sigma_{\mathbf{Y}}^{-1} \mathbf{y}.$$

Similarly, for the K_0 density, we can rewrite (56) and (60) respectively as

$$p_{K_0}(y) = \frac{1}{\sqrt{\pi}|y|\sigma_Y} e^{-2\,|y|/\sigma_Y} \tag{96}$$

$$p_{K_0}(\mathbf{y}) = \frac{\sqrt{2}+4\sqrt{s}}{2\,(\pi s)^{3/2}\ |\Sigma_{\mathbf{Y}}|}\, e^{-2\sqrt{2\,s}}. \tag{97}$$

For the Γ pdf, it is necessary to calculate the bi– and four–variates with a series expansion, as mentioned previously. It is clear from (63) and (67), however, that the functional dependence of the Γ pdf on the subband samples and their statistics enters exclusively through the terms $|\Sigma_{\mathbf{X}}|$ and $s = \mathbf{x}^T \mathbf{\Sigma}_{\mathbf{X}}^{-1} \mathbf{x}$. Hence, variates of the Γ pdf can also be specialized for complex data using the results of Theorem 1.

A.7 Partial Derivate Calculation

In order to estimate beamforming parameters with an MMI criterion using non-Gaussian pdfs, we first approximate

$$\begin{aligned} I(Y_1, Y_2) \approx \frac{1}{N} \sum_{t=0}^{N-1} \Big[&\log p\left(y_1^{(t)}, y_2^{(t)}\right) \\ &- \log p\left(y_1^{(t)}\right) - \log p\left(y_2^{(t)}\right)\Big] \end{aligned} \tag{98}$$

where

$$y_i^{(t)} = (\mathbf{w}_{\mathrm{q},i} - \mathbf{B}_i \mathbf{w}_{\mathrm{a},i})^H \mathbf{x}^{(t)}$$

for each $\mathbf{x}^{(t)}$ drawn from a *training set* $\mathcal{X} = \left\{\mathbf{x}^{(t)}\right\}_{t=0}^{N-1}$. From (98), it follows that

$$\begin{aligned} \frac{\partial I(Y_1, Y_2)}{\partial \mathbf{w}_{\mathrm{a},i}{}^*} \approx \frac{1}{N} \sum_{t=0}^{N-1} \Bigg[&\frac{\partial \log p\left(y_1^{(t)}, y_2^{(t)}\right)}{\partial \mathbf{w}_{\mathrm{a},i}{}^*} \\ &- \frac{\partial \log p\left(y_1^{(t)}\right)}{\partial \mathbf{w}_{\mathrm{a},i}{}^*} - \frac{\log p\left(y_2^{(t)}\right)}{\partial \mathbf{w}_{\mathrm{a},i}{}^*}\Bigg]. \end{aligned} \tag{99}$$

The partial derivative (99) is specialized for the Laplace, K_0 and, Γ pdfs in [16].

References

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

[2] L. C. Parra and C. V. Alvino, “Geometric source separation: Merging convolutive source separation with geometric beamforming,” *IEEE Trans. Speech Audio Proc.*, vol. 10, no. 6, pp. 352–362, Sept. 2002.

[3] 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.

[4] A. Hyvärinen and E. Oja, “Independent component analysis: Algorithms and applications,” *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[5] 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. ASRU*, 2005, pp. 357–362.

[6] Y. L. Luke, *The Special Functions and their Approximations.* New York: Academic Press, 1969.

[7] R. G. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968.

[8] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis.* New York: Wiley, 1984.

[9] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* New York: Academic Press, 1990.

[10] P. P. Vaidyanathan, *Multirate Systems and Filter Banks.* Englewood Cliffs, New Jersey: Prentice Hall, 1993.

[11] J. M. de Haan, N. Grbic, I. Claesson, and S. E. Nordholm, "Filter bank design for subband adaptive microphone arrays," *IEEE Trans. Speech Audio Proc.*, vol. 11, no. 1, pp. 14–23, Jan. 2003.

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

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

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

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

[16] J. McDonough and K. Kumatani, "Minimum mutual information beamforming," Interactive Systems Lab, Universität Karlsruhe, Tech. Rep. 107, August 2006.

[17] 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 *in Proc. Interspeech*, 2006, pp. 2594–2597.

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

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

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

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

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

[23] J. Fransen, D. Pye, T. Robinson, P. Woodland, and S. Young, "WSJCAM0 corpus and recording description," Cambridge University Engineering Department (CUED), Speech Group, Trumpington Street, Cambridge CB2 1PZ, UK, Tech. Rep. CUED/F-INFENG/TR.192, Sept. 1994.

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

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

[26] L. Uebel and P. Woodland, “Improvements in linear transform based speaker adaptation,” in *Proc. ICASSP*, vol. 1, 2001, pp. 49–52.

[27] 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.

[28] M. J. F. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, Apr. 1998.

[29] 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, Apr. 1995.

[30] S. Wolfram, *The Mathematica Book*. Cambridge: Cambridge University Press, 1996.

[31] W. Rudin, *Real and Complex Analysis*. Boston, Massachusetts: McGraw-Hill, 1987.

[32] R. E. Greene and S. G. Krantz, *Function Theory of One Complex Variable*. New York: Wiley-Interscience, 1997.

[33] T. Kailath, *Linear Systems*. Englewood Cliffs, NJ: Prentice Hall, 1980.