MINIMUM MUTUAL INFORMATION BEAMFORMING FOR SIMULTANEOUS ACTIVE SPEAKERS

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

Institute for Intelligent Sensor-Actuator Systems
University of Karlsruhe
Kaiserstrasse 12
D-76128 Karlsruhe, Germany

ABSTRACT

In this work, we address 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 adjust the* active weight vectors *of both GSCs to obtain two output signals with* minimum mutual information *(MMI). In order to calculate the mutual information of the complex subband snapshots, we consider four probability density functions (pdfs), namely the Gaussian, Laplace, K_0 and Γ pdfs. The latter three belong to the class of* super-Gaussian *density functions that are typically used in* independent component analysis *as opposed to conventional beamforming. The proposed algorithms provide effective nulling of the undesired source, but without the signal cancellation problems seen in conventional beamforming. 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. *In the experiments, the delay-and-sum beamformer achieved a word error rate (WER) of 70.4%. The MMI beamformer under a Gaussian assumption achieved 55.2% WER which was further reduced to 52.0% with a K_0 pdf, whereas the WER for data recorded with close-talking microphone was 21.6%.*

1. INTRODUCTION

One of the difficulties of recognizing speech in realistic environments is that multiple speakers talk simultaneously. It was identified that 50% of speech segments in a meeting or telephone conversation contained overlapping speech [1]. Current solutions to separate mixed speech can be classified into two techniques, acoustic beamforming and blind source separation (BSS). Generally, in BSS based techniques, the use of many microphones makes the estimation of an unmixing matrix difficult because many initial values must be specified. Thus if many microphones are available, acoustic beamforming would seem to be the more promising approach. Hence, in this work we construct two subband domain beamformers, one for each active source.

In acoustic beamforming, it is typically assumed that the position of the speaker is estimated by a speaker localization system. A conventional beamformer in *generalized sidelobe canceller* (GSC) configuration is structured such that the direct signal from the speaker is undistorted [2, §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 an undesirable *signal cancellation*. To avoid the latter, the adaptation of the active weight vectors is typically halted whenever the desired source is active.

In this work, we consider a situation where two speakers are simultaneously active. We construct one subband domain beamformer 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). Parra and Alvino [3] proposed a *geometric source separation* (GSS) algorithm with similarities to the algorithm proposed here. Their algorithm attempts to decorrelate the outputs of two beamformers. We discuss Parra and Alvino's GSS algorithm in Section 3.2, and propose novel algorithms which assume the probability density function (pdf) of subband snapshots are Gaussian and super-Gaussian.

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). As this data was recorded from actual speakers in a real, reverberant room, it provides the possibility of conducting source separation experiments under realistic conditions, which is noteably different from the vast majority of the experiments reported in the beamforming and blind source separation literature.

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 [3]. 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 conducted on data from the PASCAL Speech Separation Challenge; see Lincoln *et al.* [4] for a description of the data collection apparatus. Finally, in Section 6, we present our conclusions and plans for future work.

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* [5] of Y_1 and Y_2 can be expressed as

$$I(Y_1,Y_2) = \mathscr{E}\left\{\log\frac{p(Y_1,Y_2)}{p(Y_1)p(Y_2)}\right\} \tag{2.1}$$

where $\mathscr{E}\{\}$ indicates the ensemble expectation.

The univariate Gaussian pdf for complex r.v.s Y_i can be expressed as

$$p(Y_i) = \frac{1}{\pi\sigma_i^2}\exp\left(-|Y_i|^2/\sigma_i^2\right) \tag{2.2}$$

This work was supported by the European Union (EU) under the integrated projects AMIDA, *Augmented Multi-party Interaction with Distance Access*, contract number IST-033812 and CHIL, *Computers in the Human Interaction Loop*, contract number 506909, as well as the German Ministry of Research and Technology (BMBF) under the *SmartWeb* project, grant number 01IMD01A. The authors gratefully thank the EU and the Republic of Germany for their financial support, and all project partners for a fruitful collaboration. Kenichi Kumatani is with the Institute for Computer Science and Engineering, Intelligent Sensor-Actuator Systems (ISAS) at the University of Karlsruhe in Karlsruhe, Germany and with the IDIAP Research Institute in Martigny, Switzerland. Tobias Gehrig, Uwe Mayer, Emilian Stoimenov, and Matthias Wölfel are with the Institute for Theoretical Computer Science at the University of Karlsruhe. John McDonough is with ISAS at the University of Karlsruhe and with Spoken Language Systems at Saarland University in Saarbrücken, Germany.

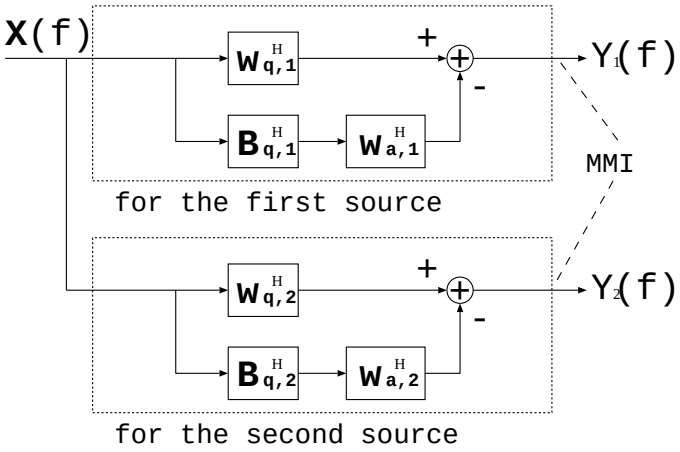

Fig. 1. A beamformer in GSC configuration.

where $\sigma_i^2 = \mathscr{E}\{Y_i Y_i^*\}$ is the variance of Y_i. Let us define the zero-mean complex random vector $\mathbf{Y} = [Y_1 \quad Y_2]^T$ and the *covariance matrix*.

$$\Sigma_Y = \mathscr{E}\{\mathbf{Y}\mathbf{Y}^H\} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} \\ \sigma_1\sigma_2\rho_{12} & \sigma_2^2 \end{bmatrix} \tag{2.3}$$

where

$$\rho_{12} = \frac{\varepsilon_{12}}{\sigma_1\,\sigma_2} \text{ and } \varepsilon_{12} = \mathscr{E}\{Y_1 Y_2^*\}$$

The bivariate Gaussian pdf for complex r.v.s is given by

$$p(Y_1, Y_2) = \frac{1}{\pi^2 |\Sigma_Y|} \exp\left(-\mathbf{Y}^T \Sigma_Y^{-1} \mathbf{Y}\right) \tag{2.4}$$

It follows that the mutual information (2.1) for jointly Gaussian complex r.v.s can be expressed as [6]

$$I(Y_1, Y_2) = -\tfrac{1}{2}\log\left(1 - |\rho_{12}|^2\right) \tag{2.5}$$

From (2.5), it is clear that minimizing the mutual information between two zero-mean Gaussian r.v.s is equivalent to minimizing the magnitude of their *cross correlation coefficient* ρ_{12}, and that $I(Y_1, Y_2) = 0$ if and only if $|\rho_{12}| = 0$.

3. BEAMFORMING

Here we formulate the MMI beamformer and compare it to the GSS algorithm of Parra and Alvino [3].

3.1 Minimum Mutual Information Beamformer

Consider two subband beamformers in GSC configuration as shown in Figure 1. The output of the *i*-th beamformer for a given subband can be expressed as,

$$Y_i = \left(\mathbf{w}_{q,i} - \mathbf{B}_i \mathbf{w}_{a,i}\right)^H \mathbf{X} \tag{3.1}$$

where $\mathbf{w}_{q,i}$ is the *quiescent weight vector* for the *i*-th source, $\mathbf{B}_i$ is the *blocking matrix*, $\mathbf{w}_{a,i}$ is the *active weight vector*, and $\mathbf{X}$ is the input subband *snapshot vector*. In keeping with the GSC formalism, $\mathbf{w}_{q,i}$ is chosen to preserve a signal from the *look direction* and, at the same time, to suppress an interference [2, §6.3]. The blocking matrix $\mathbf{B}_i$ is chosen such that $\mathbf{B}_i^H \mathbf{w}_{q,i} = \mathbf{0}$. The active weight vector $\mathbf{w}_{a,i}$ is typically chosen to maximize the signal-to-noise ratio (SNR). Here, however, we develop an optimization procedure to find that $\mathbf{w}_{a,i}$ which *minimizes* the mutual information $I(Y_1, Y_2)$. Minimizing a mutual information criterion yields a weight vector $\mathbf{w}_{a,i}$ capable of canceling interference that leaks through the sidelobes without the signal cancellation problems encountered in conventional beamforming.

The subband analysis and resynthesis can be performed with a *perfect reconstruction filterbank* such as the popular *cosine modulated filterbank* [7, §8]. Beamforming in the subband domain has the considerable advantage that the active sensor weights can be optimized for each subband independently, which saves a tremendous computation. In addition, the GSC constraint solves the problems with source permutation and scaling ambiguity typically encountered in conventional blind source separation algorithms [8].

The details of the estimation of the optimal active weights $\mathbf{w}_{a,i}$ under the MMI criterion (2.5) as well as the application of a *regularization term* are described in McDonough *et al* [6].

3.2 Geometric Source Separation

Parra and Alvino [3] proposed a *geometric source separation* (GSS) algorithm which has many similarities to the algorithm proposed. Instead of minimizing the mutual information between two signals, Parra and Alvino sought to diagonalize the cross-power spectra under geometric constraints which are equivalent to the distortionless constraint inherent in the GSC. In the case of a Gaussian pdf, the principal difference between GSS and the algorithm proposed here, is that GSS seeks to minimize $|\varepsilon_{12}|^2$ instead of $|\rho_{12}|^2$. Although the difference between minimizing $|\varepsilon_{12}|^2$ 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 $|\varepsilon_{12}|^2$, which also comprises the *numerator* of $|\rho_{12}|^2$. Such null steering is also observed in conventional adaptive beamformers [2, §6.3]. The difference between the two optimization criteria is due to the presence of the terms σ_i^2 in the denominator of $|\rho_{12}|^2$, which indicate that, in addition to nulling out the unwanted signal, improvements are possible by *increasing* the strength of the desired signal. In 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 all such strong reflections. 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 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 $\mathbf{w}_a$. 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. It is worth mentioning that the proposed criteria are completely different from GSS algothim if a super-Gaussian assumption is used.

In order to verify the behavior of the MMI beamformer described above, we plotted beam patterns for a simulated acoustic environment. As shown in Figure 2, we considered a simple configuration in which 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. Denoting the signals from each source in Figure 2 as $s_1(t)$ and $s_2(t)$, the signal $x_i(t)$ impinging on the *i*-th microphone can be expressed as

$$\begin{aligned} x_i(t) &= A_1\, s_1(t - i\cdot d\sin\theta_1) + A_2\, s_2(t - i\cdot d\sin\theta_2) \\ &+ A_3\, s_1\left(t - i\cdot d\sin\theta_3 - (\sqrt{3}-1)D/c\right) \\ &+ A_4\, s_2\left(t - i\cdot d\sin\theta_4 - (\sqrt{7}-1)D/c\right) \end{aligned}$$

where c is the speed of sound, A_n is an attenuation factor, d is the distance between the adjacent microphones, D is the distance between each souce and the center of the microphone array, and θ_n represents the direction of arrival for each wave.

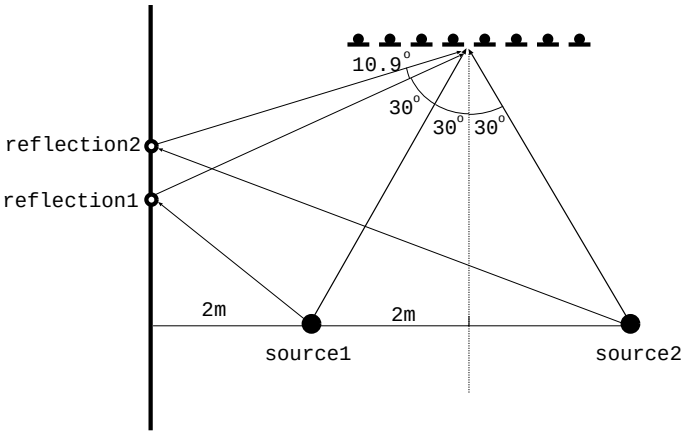

Fig. 2. Configuration of sources, sensors, and reflective surface for simulation comparing GSS and MMI beamformer.

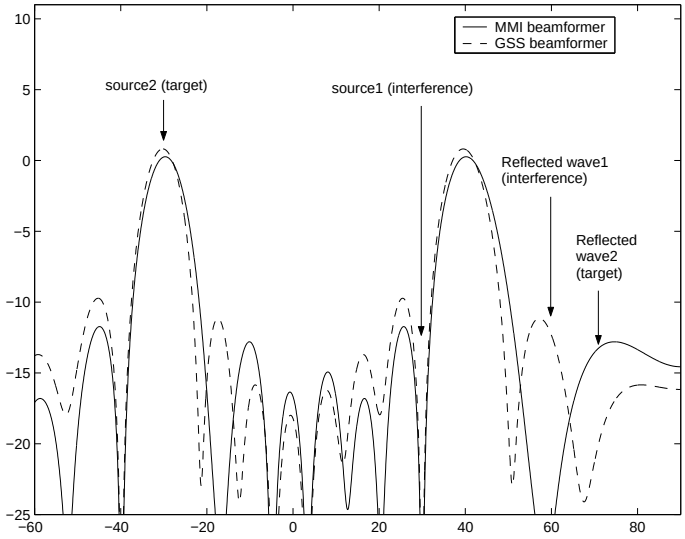

Fig. 3. Beam patterns produced by GSS and MMI beamformer.

Speech data were used as sound sources in this simulation. Figure 3 shows beam patterns at 3000 Hz obtained with the MMI beamformer and the GSS algorithm. The beam patterns formed by both techniques were such that source2 in Figure 2 was enhanced while the direct wave from source1 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 reflected signals, however, are treated very differently by the two algorithms. The MMI beamformer places a deep null on the reflection1 from the *unwanted* source and positions a sidelobe on reflection2 from the *desired* source, exactly as we would expect based on the argument above. The GSS algorithm, on the other hand, does the exact opposite, namely, it emphasizes the undesired reflection1 and suppresses the desired reflection2.

4. SUPER-GAUSSIAN PROBABILITY DENSITY FUNCTIONS

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 [5]. 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 the 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* [9]. It is well known that a Gaussian r.v. has the highest entropy of all r.v.s with a given variance [9, Thm. 7.4.1], which holds also for complex Gaussian r.v.s [10, 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.

pdf	$\frac{1}{T}\sum_{t=0}^{T-1}\log p(X_t;\text{pdf})$
Γ	-0.779
K_0	-1.11
Laplace	-2.48
Gaussian	-9.93

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

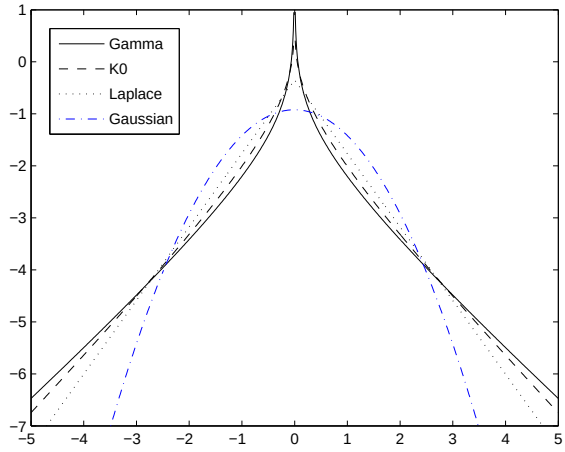

Fig. 4. Plot of the log-likelihood of the super-Gaussian and Gaussian pdfs.

Table 1 shows the average log-likelihood of subband samples of speech recorded with a close-talking microphone (CTM) as calculated with the Gaussian and three super-Gaussian pdfs, namely, the Laplace, K_0 and Γ pdfs. It is clear from these log-likelihood values that the complex subband samples of speech are in fact better modelled 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.

A plot of the log-likelihood of the Gaussian and three super-Gaussian *real* univariate pdfs considered here is provided in Figure 4. 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

$$\text{kurt}(Y) = \mathscr{E}\{Y^4\} - 3(\mathscr{E}\{Y^2\})^2$$

is a measure of how *non-Gaussian* it is [5]. 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 Figure 4, 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 improves 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 [11], Laplace, K_0 and Γ density pdfs belong to the class of *spherically invariant random processes* (SIRPs), which is a very attractive feature for two reasons. Firstly, it implies that multivariate pdfs of all orders can be readily derived from the theory of *Meijer G-functions* [12] 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

$$p_{\text{Lap}}(Y_i) = \frac{4}{\sqrt{\pi}\sigma_Y^2} K_0\left(\frac{2\sqrt{2}|Y_i|}{\sigma_Y}\right) \tag{4.1}$$

where $K_0(z)$ is the modified Bessel function of the second kind [13, §3.2.10] and $\sigma_Y^2 = \mathscr{E}\{|Y_i|^2\}$. For $\mathbf{Y} \in \mathbf{C}^2$, the bivariate Laplace pdf is given by

$$p_{\text{Lap}}(\mathbf{Y}) = \frac{16}{\pi^{3/2}|\Sigma_{\mathbf{Y}}|\sqrt{s}} K_1\left(4\sqrt{s}\right) \tag{4.2}$$

where

$$\Sigma_{\mathbf{Y}} = \mathscr{E}\{\mathbf{Y}\mathbf{Y}^H\} \text{ and } s = \mathbf{Y}^H\Sigma_{\mathbf{Y}}^{-1}\mathbf{Y}$$

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

$$p_{K_0}(Y_i) = \frac{1}{\sqrt{\pi}\sigma_Y|Y_i|} \exp\left(-2\,|Y_i|/\sigma_Y\right) \tag{4.3}$$

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

$$p_{K_0}(\mathbf{Y}) = \frac{\sqrt{2}+4\sqrt{s}}{2\pi^{3/2}\,|\Sigma_{\mathbf{Y}}|\,s^{3/2}} \exp\left(-2\sqrt{2s}\right) \tag{4.4}$$

Equations (4.1–4.4) differ from the canonical form of the real univariate Laplace and K_0 pdfs, because, unlike the Gaussian pdf, the *functional form* of a super-Gaussian pdf changes as the order of the variate is increased. Moreover, the complex univariate is actually derived from the real bivariate pdf. Similarly, the complex bivariate is derived from the real four-variate. Derivations of (4.1–4.4) are provided in [6]. For the Γ pdf, the complex univariate and bivariate pdfs *cannot* be expressed in closed form in terms of elementary or even special functions. It is possible, however, to derive Taylor series expansions that enable the required variates to be calculated to arbitrary accuracy [6].

The mutual information can no longer be expressed in closed form as in (2.5) for the super-Gaussian pdfs. We can, however, replace the exact mutual information with the *empirical mutual information*

$$I(Y_1,Y_2) \approx \frac{1}{N}\sum_{t=0}^{N-1}\left[\log p(\mathbf{Y}^{(t)}) - \sum_{i=1}^{2}\log p(Y_i^{(t)})\right] \tag{4.5}$$

Such an empirical approximation was used for the experiments described in the next section.

5. EXPERIMENTS

We performed far-field automatic speech recognition experiments on development data from the *PASCAL Speech Separation Challenge* (SSC) [4]. The data contain recordings of five pairs of speakers and each pair of speakers reads approximately 30 sentences taken from the 5,000 word vocabulary Wall Street Journal (WSJ) task. The data were 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. The database also contains speech recorded with close talking microphones (CTM). 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 convoluted with measured room impulse responses, the position of the speaker's head as well as the speaking volume varies.

Prior to beamforming, we first estimated the speaker's position with the speaker localization system described in [14]. In addition to the speaker position, our source localization system is also capable of determining when each source is active. This information proved very useful to segment the utterance of each speaker, given that the 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 from 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.

The active weights for each subband were initialized to zero for estimation with the Gaussian pdf. For estimation with the super-Gaussian pdfs, the active weights were initialized to the optimal values under the Gaussian assumption.

After beamforming, the feature extraction of our ASR system was based on cepstral features estimated with a warped *minimum variance distortionless response* [15] (MVDR) spectral envelope of model order 30. We concatenated 15 cepstral features, each of length 20, then applied linear discriminant analysis (LDA) [16, §10] and a *semi-tied covariance* (STC) [17] transform to obtain final features of length 42 for speech recognition. The far-field ASR experiments reported here were conducted entirely with the *Millenium* automatic speech recognition system. 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. The *word trace decoder* in Millenium is implemented along the lines suggested by Saon *et al.* [18], and is capable of generating word lattices, which can then be optimized with WFST operations as in [19].

The training data used for the experiments were 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 WSJCAM0 corpus [20] was used for HMM training in order to cover the British accents for the speakers [4]. Acoustic models estimated with three different HMM training schemes were used for the several decoding passes: conventional maximum likelihood (ML) HMM training [21, §12], speaker-adapted training under a ML criterion (ML-SAT) [22]. Our baseline system was fully continuous with 3,500 codebooks and a total of 180,656 Gaussian components.

We performed four passes of decoding on the waveforms obtained with each of the beamforming algorithms. Parameters for speaker adaptation were estimated using the word lattices generated during the prior pass as in [23]. A description of the individual decoding passes follows:

1. Decode with the unadapted, conventional ML acoustic model and bigram language model (LM).
2. Estimate vocal tract length normalization (VTLN) [24] parameters and constrained maximum likelihood linear regression parameters (CMLLR) [25] for each speaker, then redecode with the conventional ML acoustic model and bigram LM.
3. Estimate VTLN, CMLLR, and maximum likelihood linear regression (MLLR) [26] parameters for each speaker, then redecode with the ML-SAT model and bigram LM.
4. Estimate VTLN, CMLLR, MLLR parameters, then redecode with the ML-SAT model and bigram LM.

Table 2 shows the word error rate (WER) for every beamforming algorithm and speech recorded with the CTM after every decoding pass on the SSC data. After the fourth pass, the delay-and-sum

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

Table 2. Word error rates for every beamforming algorithm after every decoding passes.

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 MMI beamformer with a Gaussian pdf (55.2%) was somewhat better than the GSS algorithm (56.3%), which is what should be expected given the reasoning in Section 3.2. The best performance was achieved with the K_0 pdf assumption (52.0%).

Although 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 [27]. Hence, aliasing introduced by the filter bank could be masking the gain which 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 the 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 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 distortions in the output speech. We evaluated the Gaussian and three super-Gaussian pdfs in calculating the mutual information of the beamformer outputs, and found the K_0 pdf to provide the best ASR performance on the separated speech.

REFERENCES

[1] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: findings and implications for automatic processing of multi-party conversation," in *Proc. Eurospeech 2001*, 2001, vol. II.

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

[3] Lucas C. Parra and Christopher V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. Speech Audio Proc.*, vol. 10, no. 6, pp. 352–362, September 2002.

[4] M. Lincoln, I. McCowan, J. Vepa, and H.K. Maganti, "The multi-channel wall street journal audio visual corpus (mc-wsj-av): specification and initial experiments," in *Proc. ASRU*, November 2005, pp. 357–362.

[5] Aapo Hyvärinen and Erkki Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, pp. 411–430, 2000.

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

[7] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice Hall, Englewood Cliffs, 1993.

[8] 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*, pp. 255–289. Kluwer Academic, Boston, 2004.

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

[10] 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.

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

[12] Yudell L. Luke, *The Special Functions and their Approximations*, Academic Press, New York, 1969.

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

[14] T. Gehrig and J. McDonough, "Tracking and far-field speech recognition for multiple simultaneous speakers," *Proc. Workshop on Machine Learning and Multimodal Interaction*, September 2006.

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

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

[17] M. J. F. Gales, "Semi-tied covariance matrices," in *Proc. ICASSP*, 1998.

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

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

[20] Jeroen Fransen, Dave Pye, Tony Robinson, Phil Woodland, and Steve Young, "Wsjcam0 corpus and recording description," Tech. Rep. CUED/F-INFENG/TR.192, Cambridge University Engineering Department (CUED) Speech Group, September 1994.

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

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

[23] L. Uebel and P. Woodland, "Improvements in linear transform based speaker adaptation," in *Proc. ICASSP*, 2001.

[24] M. Wölfel, "Mel-Frequenzanpassung der Minimum Varianz Distortionless Response Einhüllenden," *Proc. of ESSV*, 2003.

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

[26] 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.

[27] Jan Mark de Haan, Nedelko Grbic, Ingvar Claesson, and Sven Erik Nordholm, "Filter bank design for subband adaptive microphone arrays," *IEEE Trans. Speech and Audio Proc.*, vol. 11, no. 1, pp. 14–23, January 2003.