To Separate Speech!
A System for Recognizing Simultaneous Speech

John McDonough[1,3], Kenichi Kumatani[2,3], Tobias Gehrig[4],
Emilian Stoimenov[4], Uwe Mayer[4], Stefan Schacht[1], Matthias Wölfel[4], and
Dietrich Klakow[1]

[1] Spoken Language Systems, Saarland University, Saarbrücken, Germany
[2] IDIAP Research Institute, Martigny, Switzerland
[3] Institute for Intelligent Sensor-Actuator Systems, University of Karlsruhe, Germany
[4] Institute for Theoretical Computer Science, University of Karlsruhe, Germany

Abstract. The *PASCAL Speech Separation Challenge* (SSC) is based on a corpus of sentences from the Wall Street Journal task read by two speakers simultaneously and captured with two circular eight-channel microphone arrays. This work describes our system for the recognition of such *simultaneous* speech. Our system has four principal components: A person tracker returns the locations of both active speakers, as well as segmentation information for each utterance, which are often of unequal length; two beamformers in *generalized sidelobe canceller* (GSC) configuration separate the simultaneous speech by setting their active weight vectors according to a minimum mutual information (MMI) criterion; a postfilter and binary mask operating on the outputs of the beamformers further enhance the separated speech; and finally an *automatic speech recognition* (ASR) engine based on a *weighted finite-state transducer* (WFST) returns the most likely word hypotheses for the separated streams. In addition to optimizing each of these components, we investigated the effect of the filter bank design used to perform subband analysis and synthesis during beamforming. On the SSC development data, our system achieved a word error rate of 39.6%.

1 Introduction

The *PASCAL Speech Separation Challenge* (SSC) is based on a corpus of sentences from the Wall Street Journal (WSJ) task read by two speakers simultaneously and captured with two circular eight-channel microphone arrays. This work describes our system for the automatic recognition of such *simultaneous* speech. Our system has four principal componenents: A person tracker returns the locations of both active speakers, as well as segmentation information for each utterance, which are often of unequal length; two beamformers in *generalized sidelobe canceller* (GSC) configuration separate the simultaneous speech by setting their active weight vectors according to a minimum mutual information (MMI) criterion; a postfilter and binary mask operating on the outputs of the beamformers further enhance the separated speech; and finally an *automatic speech recognition* (ASR) engine based on a *weighted finite-state transducer* (WFST) returns the most likely word hypotheses for the separated streams.

Our speaker tracking system was previously described in [1]. It is based on a *joint probabilistic data association filter* (JPDAF). The JPDAF is capable of tracking multiple targets simultaneously and consists of multiple *Kalman filters*, once for each target to be tracked [2, §6.4]. When new observations become available, they are associated either with an active target or with the *clutter model*, which models spurious acoustic events, through the calculation of posterior probabilities. After the association step, the position of each target can be updated independently through suitably modified Kalman filter update formulae.

In acoustic beamforming, it is typically assumed that the position of the speaker is estimated by a speaker localization system. A conventional beamformer in GSC configuration is structured such that the direct signal from the speaker is undistorted [3, §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.

For the speech separation task, we implemented two beamformers in GSC configuration, where one GSC was directed at each source. We then jointly adjusted the active weight vectors of the GSCs so as to provide output streams with minimum mutual information. Better speech separation was achieved through the use of non-Gaussian pdfs for calculating mutual information. In our initial experiments on the SSC development data, a simple 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%.

We also used novel techniques to represent a *full* WSJ trigram language model with 1,639,687 bigrams and 2,684,151 trigrams as a statically-expanded WFST for decoding the separated streams. The final decoding graph constructed from this trigram contained nearly 50 million states and over 100 million transitions. We were able to construct such a large decoding graph by introducing an additional symbol into the language model to explicitly model transitions to the back-off node and thereby make the language model transducer *sequential.* Because the components to be composed together to create the final decoding graph were likewise sequential, we were able to forego the last determinization step, which is usually the most demanding operation in terms of computation and main memory requirements. The use of the full trigram provided a reduction in WER from 52.5% to 47.7%.

In a final set of experiments, we used four different filter bank designs to perform subband analysis and synthesis. We also tested different postfiltering configurations, and applied binary masking to the postfiltered streams. Our best current result on the SSC development data is 39.6% WER. Our best result on the SSC 2007 evaluation set was 46.9% WER.

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 then discuss our MMI beamforming criterion and present the framework needed to apply minimum mutual information beamforming when the Gaussian assumption is relaxed. In Section 3 we describe sequence of operations used to optmize the search space for automatic recognition on the separated streams of speech. We also present the sizes of the language models and decoding graphs used for our experiments. In Section 4, we present the results of far-field automatic speech recognition experiments conducted on data from the PASCAL Speech Separation Challenge. Finally, in Section 5, we present our conclusions and plans for future work.

2 Beamforming

Consider two r.v.s Y_1 and Y_2. By definition, the *mutual information* [4] between Y_1 and Y_2 can be expressed as

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

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

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

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

where $\sigma_i^2 = \mathcal{E}\{Y_i Y_i^*\}$ is the variance of Y_i. Let us define the zero-mean complex random vector $\mathbf{Y} = \begin{bmatrix} Y_1 \; Y_2 \end{bmatrix}^T$ and the *covariance matrix.*

$$\Sigma_Y = \mathcal{E}\{\mathbf{Y}\mathbf{Y}^H\} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} \\ \sigma_1\sigma_2\rho_{12} & \sigma_2^2 \end{bmatrix} \tag{3}$$

where $\rho_{12} = \epsilon_{12}/\sigma_1\sigma_2$ and $\epsilon_{12} = \mathcal{E}\{Y_1 Y_2^*\}$. The bivariate Gaussian pdf for complex r.v.s is given by

$$p(Y_1, Y_2) = \frac{1}{\pi^2|\Sigma_Y|} \exp\left(-\mathbf{Y}^T \Sigma_Y^{-1} \mathbf{Y}\right) \tag{4}$$

It follows that the mutual information (1) for jointly Gaussian complex r.v.s can be expressed as [5]

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

From (5), it is clear that minimizing the mutual information between two zero-mean Gaussian r.v.s is equivalent to minimizing the magnitude of their *cross correlation coefficient* ρ_{12}, and that $I(Y_1, Y_2) = 0$ if and only if $|\rho_{12}| = 0$.

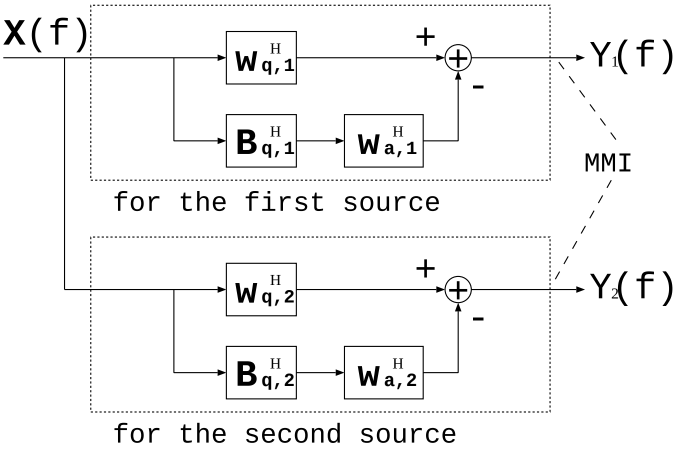

Fig. 1. A beamformer in GSC configuration.

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,

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

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* [3, §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 $I(Y_1, Y_2)$ where Y_1 and Y_2 are the outputs of the two beamformers. 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 details of the estimation of the optimal active weights $\mathbf{w}_{a,i}$ under the MMI criterion (5) as well as the application of a *regularization term* are described in Kumatani *et al* [6].

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. The subband analysis and resynthesis can be performed with a *perfect reconstruction* (PR) filter bank such as the popular *cosine modulated filter bank* [7, §8]. As this PR filter bank is based on assumptions that are not satisfied in beamforming and adptive filtering applications, however, there are other designs that are better suited for such applications. In Section 4 we present the results of ASR experiments comparing the effectiveness of frequency domain beamforming with subband beamforming based on the PR design, the design proposed by De Haan *et al* [8], and a further novel design technique. We also compare the performance of subband beamformers based on the filter designs with frequency domain beamforming based on a simple FFT.

A plot of the log-likelihood of the Gaussian and three super-Gaussian *real* univariate pdfs considered here is provided in Figure 2. 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. As explained in [6], univariate and bivariate forms of the complex Laplace, K_0 and Γ pdfs can be derived using the theory of *Meijer G-functions* [9].

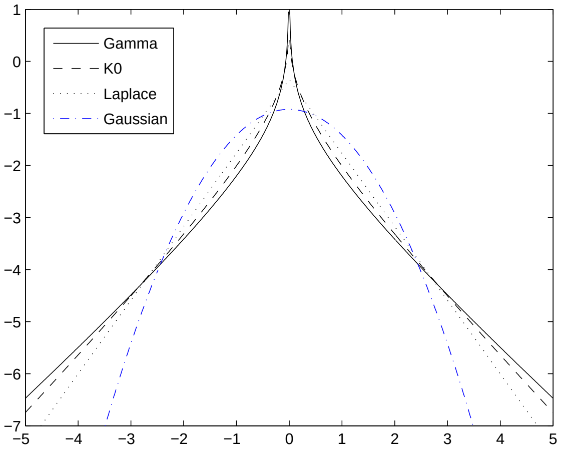

Fig. 2. Plot of the log-likelihood of the super-Gaussian and Gaussian pdfs.

3 Search Space Optimization

As originally proposed by Mohri *et al* [10, 11], a *weighted finite-state transducer* (WFST) that translates phone sequences into word sequences can be obtained by forming the *composition* $L \circ G$, where L is a *lexicon* which translates the phonetic transcription of a word to the word itself, and G is a *grammar* or *language model* which assigns to valid sequences of words a weight consisting of the negative log probability of this sequence. In the original formulation of Mohri and Riley [12], phonetic context is modeled by the series of compositions $H \circ C \circ L \circ G$, where

H is a transducer converting sequences of Gaussian mixture models (GMMs) to sequences of polyphones, and C is a transducer that converts these polyphone sequences to corresponding sequences of phones.

In [13], we showed how the necessity of explicitly modeling C could be circumvented by constructing a transducer HC that maps directly from sequences of GMM names to sequences of phones. In more recent work [14], we demonstrated that HC can be incrementally expanded and immediately determinized. Such an incremental procedure enables a much larger decision tree to be modeled as a WFST. In our previous work, we constructed a recognition network based on

$$\min \operatorname{push} \det(\min \det HC \circ \det(L \circ G)) \tag{7}$$

where det, push, and min represent the WFST equivalence transformations, *determinization* [15], *weight pushing* [16], and *minimization* [17]. The sequence represented by (7) is the "standard" build procedure [11]. By far, the most memory and time intensive portion of this build sequence is the determinization after HC has been statically composed with $L \circ G$. Hence, we sought to construct a larger recognition network by eliminating this costly determinization.

In the context of WFSTs, ϵ–symbols represent that null symbol that consumes no input or produces no ouput. A *sequential transducer* is deterministic on the input side and has no ϵ–symbols as input. A well-known theorem from Mohri [15] states that the composition of two sequential transducers is sequential. As typically constructed, the grammar G is *not* sequential, as ϵ–symbols are used to allow transitions to nodes modeling *back-off* probabilities, which in turn implies $L \circ G$ is not sequential. We remedied this problem by replacing the ϵ–symbols in G with an explicit *back-off symbol* %, much the way word end markers are introduced to disambiguate homophones [11] thereby allowing $L \circ G$ to be determinized. Similarly, a back-off self-loop was added to the end of each word in L, and to the end of each three-state sequence in HC. These changes were sufficient to make $L \circ G$ sequential. As HC was already sequential, we were able to entirely forego the determinization after composing HC and $L \circ G$. Adding the back-off symbols had an additional salutary effect in that $L \circ G$, and hence the final recognition network, became much smaller, which provided for the use of a still larger language model G.

The sizes of the shrunken and full bigram trigram language models along with the decoding graphs built from them and used for the speech recognition experiments reported in Section 4 are given in Table 1. We performed our initial experiments with the a decoding graph built based on (7) *without* explicit back-off symbols. Thereafter, we built decoding graphs with the full bigram, then with shrunken and full trigrams using the new build technique *with* explicit back-off symbols in the LM. It is worth noting that the decoding graph built from the full bigram with the back-off symbols actually has fewer nodes and transitions than the decoding graph built from the shrunken bigram without back-off symbols. Moreover, as we were able to eliminate the costly determinization after composing HC and $L \circ G$, we were able build a decoding graph from the full WSJ trigram with nearly 50 million states and over 100 million transitions.

Language	G		$HC \circ L \circ G$	
Model	Bigrams	Trigrams	Nodes	Arcs
Shrunken Bigram	323,703	0	4,974,987	16,672,798
Full Bigram	835,688	0	4,366,485	10,639,728
Shrunken Trigram	431,131	435,420	14,187,005	32,533,593
Full Trigram	1,639,687	2,684,151	49,082,515	114,304,406

Table 1. Dimensions of the various language models and the decoding graphs built from them.

4 Experiments

We performed far-field automatic speech recognition experiments on development data from the *PASCAL Speech Separation Challenge* (SSC) [18]. 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.

After beamforming, the feature extraction of our ASR system was based on cepstral features estimated with a warped *minimum variance distortionless response* [19] (MVDR) spectral envelope of model order 30. We concatenated 15 cepstral features, each of length 20, then applied linear discriminant analysis (LDA) [20, §10] and a *semi-tied covariance* (STC) [21] transform to obtain final features of length 42 for speech recognition.

4.1 Beamforming Experiments

The training data used for our initial beamforming 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 [22] was used for HMM training in order to cover the British accents for the speakers [18]. Acoustic models estimated with two different HMM training schemes were used for the several decoding passes: conventional maximum likelihood (ML) HMM training [23, §12] and speaker-adapted training under a ML criterion (ML-SAT) [24]. 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

Beamforming Algorithm	Pass (%WER)			
	1	2	3	4
Delay & Sum	85.1	77.6	72.5	70.4
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.

using the word lattices generated during the prior pass [25]. A description of the individual decoding passes follows:

1. Decode with the unadapted, conventional acoustic model and bigram language model (LM).
2. Estimate vocal tract length normalization (VTLN) [26] parameters and constrained maximum likelihood linear regression parameters (CMLLR) [27] for each speaker, then redecode with the conventional acoustic model and bigram LM.
3. Estimate VTLN, CMLLR, and maximum likelihood linear regression (MLLR) [28] parameters for each speaker, then redecode with the conventional 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 development data. These results were obtained with subband-domain beamforming where subband analysis and synthesis was performed with the perfect reconstruction cosine modulated filter bank described in [7, §8]. After the fourth pass, the delay-and-sum beamformer has the worst recognition performance of 70.4% WER. The MMI beamformer with a Gaussian achieved a WER of 55.2%. 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.

4.2 Language Modeling Experiments

To test the effect of language modeling improvements, we trained a triphone acoustic model on 30 hours of American WSJ data, and the 12 hours of Cambridge WSJ data. For the language modeling experiments, we used the same acoustic features and same sequence of decoding passes as in the prior section. The word error rate reduction achieved through larger language models are shown in Table 3. The most dramatic reduction in WER was achieved by replacing the bigram LMs with the shrunken trigram. As shown in Table 1, the

Table 3. ASR results on the SSC development data.

	Pass (%WER)			
Language Model/Pass	1	2	3	4
shrunken bigram	85.7	64.8	53.8	52.5
full bigram		65.5	53.8	52.4
shrunken trigram	86.1	61.7	49.8	47.7
full trigram	88.3	61.0	48.9	47.0

shrunken trigram produced a decoding graph that was still small enough to run on our standard 32-bit workstations. The full trigram, on the other hand, could not be used on the 32-bit machines used for the experiments reported here. On a work station with a 64-bit operating system, more than 7 Gb of RAM were required merely to load the decoding graph built from the full trigram, and the entire task image of the recognition job was approximately 8 Gb. Moreover, the reduction in WER with respect to the shrunken trigram that was achieved by the full trigram was less than one percent absolute. Hence, we used the decoding graph built from the shrunken trigram to decode the evaluation data, as that system was much more tractable.

The results of further experiments with these language models, as well as a description of a technique for dynamically composing the HC and $L \circ G$ components, and thereby radically reducing the enormous amount of random access memory required by the full trigram, are given in [29].

4.3 Filter Bank Experiments

As explained in Section 2, our MMI beamformer operates in the frequency or subband domain. Hence, the digital filter bank used for subband analysis and resynthesis is an important component of the speech separation system. We investigated four different filter bank designs, including:

1. The *cosine modulated filter bank* described by Vaidyanathan [7, §8], which yields *perfect reconstruction* (PR) under optimal conditions. In such a filter bank, PR is achieved through *aliasing cancellation*, wherein the aliasing that is perforce present in one subband is cancelled by the aliasing in all others. Aliasing cancellation breaks down if arbitrary complex factors are applied to the subband samples. For this reason, such a PR filter bank is not optimal for beamforming or adaptive filtering applications.
2. An DFT filter bank based on overlap-add.
3. The modulated filter bank proposed by De Haan *et al* [8], wherein separate analysis and synthesis prototypes are designed to minimize an error criterion consisting of a weighted combination of the total spectral response error and the aliasing distortion. This design is dependent on the use of oversampling to reduce aliasing error.

4. A novel cosine modulated design which differs from the De Haan filter bank in that a *Nyquist(M)* constraint [7, §4] is imposed on the prototype in order to ensure that the total response error vanishes. Thereafter the remaining components of the prototype are chosen to minimize aliasing error, as with the De Haan design. The Nyquist(M) design similarly uses oversampling to reduce aliasing distortion.

The word error rates (WERs) obtained with the four filter banks on the SSC development data are shown in Table 4. For these experiments, the Gaussian pdf was used exclusively. We also investigated the effect of applying a Zelinski post filter [30] to the output of the beamformer in the subband domain, as well as the binary mask [5] described in [31]. The results indicate that the performnace of PR filter bank is actually quite competitive if no postfiltering nor binary masking is applied to the output of the beamformer. For the PR design, performance *degrades* from 50.7% WER to 52.5% when such postfiltering and masking is applied, which is not surprising given that both will tend to destroy the aliasing cancellation on which this design is based. When postfiltering and masking is applied to either the De Haan or the Nyquist(M) designs, performance is greatly enhanced. With the De Haan design adding postfiltering and masking reduced WER from 53.5% to 39.6%. With postfiltering and masking the Nyquist(M) design achieved very similar performance of 40.9%. For both the De Haan and Nyquist(M) designs, an oversampling factor of eight was used. The simple FFT achieved significantly worse performance than all of the subband filter banks.

Table 4. ASR results on the SSC development data.

Filter Bank	Pass (%WER)			
	1	2	3	4
PR	87.7	65.2	54.0	50.7
PR + postfilter + binary mask	87.1	66.6	55.7	52.5
FFT	88.5	71.1	58.8	55.5
De Haan	88.7	68.2	56.1	53.3
De Haan + postfilter + binary mask	82.7	57.7	42.7	39.6
Nyquist(M) + postfilter + binary mask	84.8	58.0	43.4	40.9

5 Conclusions and Future Work

In this work, we have described our system for the automatic recognition of simultaneous speech. Our system consisted of three principal components: A

[5] We learned of the binary masking technique only by attending MLMI and listening to Iain McCowan's presentation about the SSC system developed by him and his collaborators. Our experiments with the binary mask were conducted *after* the SSC deadline.

person tracker returns the locations of both active speakers, as well as segmentation information for each utterance, which are often of unequal length; two beamformers in GSC configuration separate the simultaneous speech by setting their active weight vectors according to a minimum mutual information (MMI) criterion; a postfilter and binary mask operating on the outputs of the beamformers further enhance the separated speech; and finally an ASR engine based on a WFST returns the most likely word hypotheses for the separated streams. In addition to developing and optimizing each of these three components, we have also proposed a novel filter bank design in this work that, when used for subband beamforming, provided ASR performance comparable or superior to any design that has previously appeared in the literature. Our final results on the SSC development data were 39.6% WER. On the SSC evaluation data, our system achieved a WER of 46.2%.

In future, we plan to continue our investigations into the use of super-Gaussians pdfs for MMI beamforming. This will entail systematically searching the entire class of multi-dimensional super-Gaussians pdfs that can be represented with the aid of the Meijer-G function. We will also develop an online or LMS-style algorithm for updating the active weight vectors of the GSCs during MMI beamforming. Finally, we hope to investigate other optimization criteria such the *negentropy* metric typically used in the field of independent component analysis [4].

References

1. Tobias Gehrig, Ulrich Klee, John McDonough, Shajith Ikbal, Matthias Wölfel, and Christian Fügen, "Tracking and beamforming for multiple simultaneous speakers with probabilistic data association filters," in *in Proc. Interspeech*, 2006, pp. 2594–2597.
2. Y. Bar-Shalom and T. E. Fortmann, *Tracking and Data Association*, Academic Press, San Diego, 1988.
3. H. L. Van Trees, *Optimum Array Processing*, Wiley-Interscience, New York, 2002.
4. Aapo Hyvärinen and Erkki Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, pp. 411–430, 2000.
5. J. McDonough and K. Kumatani, "Minimum mutual information beamforming," Tech. Rep. 107, Interactive Systems Lab, Universität Karlsruhe, August 2006.
6. Kenichi Kumatani, Tobias Gehrig, Uwe Mayer, Emilian Stoimenov, John McDonough, and Matthias Wölfel, "Adaptive beamforming with a minimum mutual information criterion," *IEEE Trans. Audio Speech and Lang. Proc.*, to appear.
7. P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice Hall, Englewood Cliffs, 1993.
8. 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.
9. Helmut Brehm and Walter Stammler, "Description and generation of spherically invariant speech-model signals," *Signal Processing*, vol. 12, pp. 119–141, 1987.
10. M. Mohri, M. Riley, D. Hindle, A. Ljolje, and F. Periera, "Full expansion of context-dependent networks in large vocabulary speech recognition," in *Proc. ICASSP*, Seattle, 1998, vol. II, pp. 665–668.

11. M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech and Language*, vol. 16, pp. 69–88, 2002.
12. M. Mohri and M. Riley, "Network optimizations for large vocabulary speech recognition," *Speech Communication*, vol. 25, no. 3, 1998.
13. Emilian Stoimenov and John McDonough, "Modeling polyphone context with weighted finite-state transducers," in *Proc. ICASSP*, 2006.
14. Emilian Stoimenov and John McDonough, "Memory efficient modeling of polyphone context with weighted finite-state transducers," in *Proc. Interspeech*, 2007.
15. M. Mohri, "Finite-state transducers in language and speech processing," *Computational Linguistics*, vol. 23, no. 2, 1997.
16. M. Mohri and M. Riley, "A weight pushing algorithm for large vocabulary speech recognition," in *Proc. ASRU*, Aarlborg, Denmark, Sep. 2001, pp. 1603–1606.
17. Mehryar Mohri, "Minimization algorithms for sequential transducers," *Theoretical Computer Science*, vol. 234, no. 1–2, pp. 177–201, 2000.
18. 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.
19. 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.
20. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, New York, 1990.
21. M. J. F. Gales, "Semi-tied covariance matrices," in *Proc. ICASSP*, 1998.
22. 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.
23. J. Deller, J. Hansen, and J. Proakis, *Discrete-Time Processing of Speech Signals*, Macmillan Publishing, New York, 1993.
24. T. Anastasakos, J. McDonough, R. Schwarz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP*, 1996, pp. 1137–1140.
25. L. Uebel and P. Woodland, "Improvements in linear transform based speaker adaptation," in *Proc. ICASSP*, 2001.
26. M. Wölfel, "Mel-Frequenzanpassung der Minimum Varianz Distortionless Response Einhüllenden," *Proc. of ESSV*, 2003.
27. M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, 1998.
28. 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.
29. John McDonough, Emilian Stoimenov, and Dietrich Klakow, "An algorithm for fast composition of weighted finite-state transducers," in *Proc. ASRU*, submitted, 2007.
30. K. Uwe Simmer, Joerg Bitzer, and Claude Marro, "Post-filtering techniques," in *Microphone Arrays*, M. Branstein and D. Ward, Eds., pp. 39–60. Springer, Heidelberg, 2001.
31. I. McCowan, M. Hari-Krishna, D. Gatica-Perez, D. Moore, and S. Ba, "Speech acquisition in meetings with an audio-visual sensor array," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, July 2005.