ON HIDDEN MARKOV MODEL MAXIMUM NEGENTROPY BEAMFORMING

Barbara Rauch, Kenichi Kumatani, Friedrich Faubel, John McDonough, and Dietrich Klakow

Spoken Language Systems
Saarland University
D66123 Saarbrücken, Germany
{barbara.rauch,kenichi.kumatani,friedrich.faubel,
john.mcdonough,dietrich.klakow}@lsv.uni-saarland.de

ABSTRACT

In prior work, we developed a beamforming algorithm intended for automatic recognition of speech data captured with an array of distant microphones. In addition to enforcing a distortionless contraint in a desired direction, we adjusted the sensor weights so as to maximimize a *negentropy* criterion. Negentropy is a measure of how *non-Gaussian* the probability density function (pdf) of a random variable is. It is known that subband samples of speech are highly non-Gaussian, but become more Gaussian when corrupted with noise or reverberation. Here we extend our prior algorithm by using an auxiliary hidden Markov model to model the *non-stationarity* of speech during beamforming. In a set of far-field ASR experiments on data from the Multi-Channel Wall Street Journal Audio-Visual Corpus, we were able to reduce the word error rate from 14.6% to 13.6% by accounting for this non-stationarity.

***Index Terms*—** microphone arrays, beamforming, speech recognition, speech enhancement, adaptive processing

1. INTRODUCTION

With the increased availability of multi-channel audio processing hardware, the use of array processing techniques to enhance speech captured with far-field sensors prior to automatic recognition is becoming more common. The development of such novel *beamforming* algorithms is the focus of a great deal of current research. Among the approaches that have recently appeared in the literature was that of Seltzer *et al.* [1], wherein the sensor weights of an adaptive beamformer were adjusted to maximize the likelihood, as measured with a hidden Markov model (HMM), of the cepstral features extracted from the beamformer output. In addition to this, several variations on the conventional *minimum variance distortionless response* (MVDR) beamformer have been tried, including the *robust generalized sidelobe canceller* (RGSC). The RGSC based approaches could be classified into the following:

1. updating the *active weight vector* only when noise signals are dominant [2],
2. blocking the leakage of desired signal components into the sidelobe canceller by designing the *blocking matrix* [2], and
3. using acoustic transfer functions from a desired source to microphones instead of just compensating time delays [3].

The work presented here extends the *maximum negentropy beamformer* (MNB) [4], which is based on the observation that the distribution of clean speech is more non-Gaussian than that of mixture signals like noisy or reverberant speech. As negentropy is a measure of non-gaussianity, maximizing the negentropy of the beamformer output can be expected to produce an enhanced signal that provides a good estimate of the original speech of the desired source. Indeed, such an optimization criterion has proven to achieve far-field ASR performance superior to that obtained with a conventional minimum mean-squared-error beamformer [5].

There are three properties of any given signal that can potentially be exploited for beamforming or source separation, apart from the geometric information that comes with knowledge of the geometry of the sensor array as well as the position of the desired source or sources: non-whiteness, non-Gaussianity, and non-stationarity [6]. The MNB exploits the first two of these three characteristics, but it ignores the last. The extension presented in this work aims to account for the non-stationarity of speech, and does so through the use of HMMs. A first HMM decoding pass can give us information on phoneme boundaries, and this, together with the cepstral mean estimates accumulated during the training of the HMMs, provides a way to obtain a robust estimate of the statistics needed to calculate the negentropy, specifically the variance of the subband samples.

We evaluate our extension, the *hidden Markov model maximum negentropy beamformer* (HMM-MNB), on a real dataset of far-field speech with a state-of-the-art ASR system. Initial results demonstrate improvements in word error rate over a baseline which computes the variance directly from the test utterance.

The remainder of this paper is structured as follows. Sections 2 and 3 review the MNB and the Generalized Gaussian pdf, which we use to model the subband samples, respectively. The details of the HMM-MNB are described in section 4, and section 5 describes the ASR experiments and discusses the results. Our conclusions and plans for further work given in section 6.

2. MAXIMUM NEGENTROPY BEAMFORMING

Let us denote the kth subband sample produced by a beamformer in generalized sidelobe configuration as

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

where $\mathbf{w}_\text{q}$ is the quiescent vector set to satisfy a distortionless constraint in the look direction, the *blocking matrix* $\mathbf{B}$ satisfies $\mathbf{B}^H\mathbf{w}_\text{q} = \mathbf{0}$, the *active weight vector* $\mathbf{w}_\text{a}$ is chosen to optimize a statistical criterion, and $\mathbf{X}$ is the *subband domain snapshot* present at the input of the beamformer. The MNB [5] takes as optimization criterion the *differential negentropy* which, by definition, is given by the expected likelihood ratio,

$$J(Y) \triangleq E\left\{\log \frac{p(Y)}{p_\text{Gauss}(Y)}\right\}, \tag{1}$$

where $p_\text{Gauss}(Y)$ and $p(Y)$ are, respectively, Gaussian and non-Gaussian pdfs. The basic negentropy criterion can be augmented

The authors are grateful for the financial support of the German Research Foundation (DFG) in connection with the international research training network IRTG 715 “Language Technology and Cognitive Systems”.

with a regularization term intended to add robustness by penalizing large active weight vectors, according to

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

for some real $\alpha > 0$. Both $p_{\text{Gauss}}(Y)$ and $p_{\text{gg}}(Y)$ can be modeled with the circular complex pdfs discussed in section 3. In addition to beamforming, the speech can be further enhanced through, for example, Zelinski post-filtering [7].

3. GENERALIZED GAUSSIAN PDF

Kumatani *et al.* [5] present theoretical arguments and empirical evidence that subband samples of speech are not Gaussian-distributed but can be modelled well by the *generalized Gaussian* (GG) pdf. We follow their approach in using the GG pdf in the calculation of the negentropy of the beamformer outputs. The GG pdf is specified by three free parameters, a mean, a *scale factor* $\hat{\sigma}$, and a *shape factor* f. The GG pdf for a zero-mean, circular, *complex* variable z can be expressed as

$$p_{\text{gg}}(z) \triangleq \frac{f}{2\pi\, \hat{\sigma}^2\, B_{\text{c}}^2(f)\, \Gamma(2/f)} \exp\left\{ - \left| \frac{z}{\hat{\sigma}\, B_{\text{c}}(f)} \right|^f \right\}. \tag{3}$$

The normalization term

$$B_{\text{c}}(f) \triangleq \left[\frac{\Gamma(2/f)}{\Gamma(4/f)} \right]^{1/2} \tag{4}$$

ensures that the square of the scale factor is equal to the variance. In the present work, the pdf (3) is used to model the complex subband samples at the output of the beamformer.

In prior work [4], the moment and maximum likelihood (ML) methods [8, 9] were used to estimate the scale factor $\hat{\sigma}$ and shape parameter f of the GG pdf from training data. The shape parameter was held fixed during actual beamforming, and the scale factor was estimated as the square root of the global variance for the given utterance calculated from the beamforming output in each optimization iteration. Such an approach is suboptimal in that a global variance estimate disregards the non-stationary, short-term variations of human speech. In this work, we derive a time-dependent estimate of the scale factor from an auxiliary HMM as described in the next section. Given such estimates for some training data, the shape factor is then determined according to a maximum likelihood criterion.

4. HIDDEN MARKOV MODEL MAXIMUM NEGENTROPY BEAMFORMING

This work aims to account for the non-stationarity of human speech through the introduction of time-dependent scale factor estimates $\hat{\sigma}_m(k)$, where k is a time index. While it would be possible to estimate $\hat{\sigma}_m(k)$ over a shorter window, this could prove problematic. If the window is short, the estimate may not be robust, and if it is too long, we ignore the local phone structure inherent in speech. Since HMMs model that structure and are trained on sufficient data, we might expect a useful estimate of the variance if we can derive a frame-dependent variance value from the statistics stored in the HMM. Before we describe how we use the cepstral mean stored in the Gaussian mixture model of an HMM state in this way, let us describe the overall procedure we apply to optimize the active weights with such HMM-based variance estimates for a given test utterance:

1. An initial beamforming step combines the different channels into a single one. In the experiments of this study we initialized the weights to those obtained with the MNB.
2. We then use our speaker-adapted ASR system to decode the single channel test utterance and obtain an alignment between frames and HMM states / Gaussian mixtures.
3. For each frame, the variance is estimated by reconstructing the *power spectral density* (PSD) from the cepstral mean associated with a state in an *auxiliary* HMM. The auxiliary HMM is trained solely with fixed cepstral components (i.e., no delta and delta-delta components), and without Mel filters. The actual alignment of frames to states is done with the full HMM used for ASR. The Viterbi alignments can be transfered between the two HMMs because both are based on the context-dependency information.
4. For each bin, the active weights are determined that maximize the negentropy over all frames. In the calculation of the negentropy the scale parameter comes from the last step and the fixed shape parameter was estimated previously on training data. We use a conjugate gradient method to search for the optimal weights. The calculation of the gradient, which needs to account for a changing cepstral mean, is presented in Section 4.2.

4.1. Reconstruction of the Power Spectral Density

For our negentropy calculations we need an estimate of the variance in the subband domain. The first point we note is that if we know the mean of the power spectrum, we have an estimate of the subband variance $\sigma_m^2 = \mathbb{E}\{|Y_m|^2\}$ where $Y_m = Y(\omega_m)$ is the mth subband sample obtained from the analysis bank of a uniform DFT filter bank, the design of which is described in [10, §3.4]. This is because the mean value of the PSD is the average of the square of the subband magnitude. Our aim is therefore to obtain the mean PSD value. We will now show the relationship between the mean cepstral vector and the mean PSD vector.

Let $\mathbf{Y}(k)$ and $\mathbf{c}(k)$ denote the kth vectors of subband samples and cepstral coefficients, respectively, where k as an index over time. Then the relationship between $\mathbf{Y}(k)$ and $\mathbf{c}(k)$ can be expressed as

$$\mathbf{c}(k) = \mathbf{T}_\nu \log |\mathbf{Y}(k)|^2 \tag{5}$$

where $\mathbf{T}_\nu$ is the Type 2 *discrete cosine transform* (DCT) matrix which has been truncated to ν *rows*. In (5), the square magnitude and logarithm are calculated individually for each component $Y_m(k)$ of $\mathbf{Y}(k)$. Typically, ν will assume a relatively small value (e.g., $\nu = 13$), which implies it will model only the spectral envelope due to the resonances of the vocal tract. The more rapid variations due to the harmonic structure of voiced speech must then be modeled by the GG pdfs. As mentioned previously, no Mel warping is applied to the subband samples prior to the DCT, as this would only decrease the frequency resolution of the filter bank, thereby leading to suboptimal beamforming performance. The inverse $\mathbf{T}^{-1}$ of the Type 2 DCT matrix is equivalent to a Type 3 DCT matrix whose components have all been scaled by a factor of $2/M$ where M is the number of subbands used for beamforming.

If we calculate the cepstral mean $\bar{\boldsymbol{\mu}}$ over K frames, we have

$$\bar{\boldsymbol{\mu}} \triangleq \frac{1}{K} \sum_{k=0}^{K-1} \mathbf{c}(k) = \frac{1}{K} \mathbf{T}_\nu \sum_{k=0}^{K-1} \log |\mathbf{Y}(k)|^2 \,. \tag{6}$$

Now let

$$\hat{\boldsymbol{\mu}}(k) \triangleq \mathbf{A}^{(s)} \boldsymbol{\mu}(k) + \mathbf{b}^{(s)}$$

denote the transformed speaker-dependent mean aligned to the kth frame of subband samples by the Viterbi algorithm, where $\mathbf{A}^{(s)}$ and $\mathbf{b}^{(s)}$ are, respectively, a transformation matrix and cepstral bias vector intended to compensate for the unique characteristics of the voice of speaker s. For the experiments reported in this work, $\mathbf{A}^{(s)}$ and $\mathbf{b}^{(s)}$ were determined from a sparsely parameterized *all-pass transforms* [11]. Let us further define

$$\tilde{\boldsymbol{\mu}}(k) \triangleq \hat{\boldsymbol{\mu}}(k) + \bar{\boldsymbol{\mu}}, \tag{7}$$

which implies that the cepstral mean $\bar{\boldsymbol{\mu}}$ for a given utterance must be added back to the transformed speaker-dependent mean $\hat{\boldsymbol{\mu}}(k)$ in order to obtain the *true* mean $\tilde{\boldsymbol{\mu}}(k)$ for the kth cepstral frame. This is necessary because the cepstral mean was originally subtracted off during feature extraction to normalize for short term channel effects.

The diagonal covariance matrix $\boldsymbol{\Sigma}_{\mathbf{Y}}(k)$ of the kth frame of subband components can be approximated as

$$\boldsymbol{\Sigma}_{\mathbf{Y}}(k) \approx \exp(\boldsymbol{\eta}(k)), \tag{8}$$

where

$$\boldsymbol{\eta}(k) \triangleq \mathbf{T}_{\nu}^{-1}\tilde{\boldsymbol{\mu}}(k) = \mathbf{T}_{\nu}^{-1}\hat{\boldsymbol{\mu}}(k) + \bar{\boldsymbol{\eta}}, \tag{9}$$

$$\bar{\boldsymbol{\eta}} = \frac{1}{K}\mathbf{T}_{\nu}^{-1}\mathbf{T}_{\nu}\sum_{k'=0}^{K-1}\log|\mathbf{Y}(k')|^2, \tag{10}$$

and $\mathbf{T}_{\nu}^{-1}$ denotes the inverse of $\mathbf{T}$ truncated to ν *columns*. As with the square magnitude and logarithm, the exponential operation in (8) is applied component by component. Clearly $\boldsymbol{\Sigma}_{\mathbf{Y}}(k)$ is the desired power spectral density.

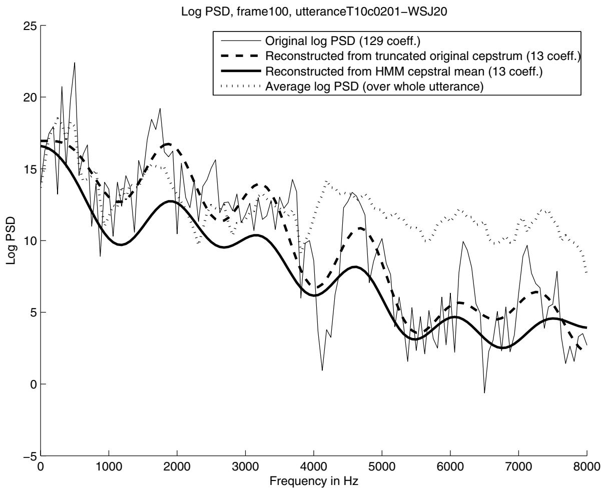

Fig. 1. *Original and reconstructed log PSD of a test utterance.*

Figure 1 shows an example of the reconstructed log PSD for one frame of a test utterance beamformed with the MNB. The average PSD value is compared to the original PSD and the PSD reconstructed from the 13 original cepstral coefficients, as well as that reconstructed from the 13-coefficient cepstral mean in the HMM state aligned with this frame. We observe that the HMM-based reconstruction approximates the spectral envelope well in the log PSD domain. The PSD obtained by averaging over the entire utterance, on the other hand, only approximates the long-term *spectral tilt*.

In earlier work [4], time averages of the moments of $Y_m(k)$ were used to calculate the differential entropies required to evaluate (1). For our initial studies on HMM negentropy beamforming, we chose instead to replace the exact differential entropy with the *empirical negentropy*, which can be expressed as

$$J_{\rm e}(Y) \triangleq \frac{1}{K}\sum_{k=0}^{K-1}\left[\log\frac{p_{\rm gg}(Y(k))}{p_{\rm Gauss}(Y(k))}\right] - \alpha\|\mathbf{w}_{\rm a}\|^2, \tag{11}$$

where $\alpha\|\mathbf{w}_{\rm a}\|^2$ is once more a regularization term.

Given this definition, our baseline will be the calculation of the negentropy based on the variance of the beamformer output computed over the complete utterance, using the empirical entropy. We refer to this baseline as the global variance case.

Since the first decoding pass will not always give correct results, we also provide results for an oracle experiment with optimistic HMM alignments, that is alignments obtained with the correct transcriptions.

4.2. Calculation of the Gradient

During beamforming we determine the active weights that maximize the optimization criterion (11). We now derive the gradient expression used by the conugate gradient optimization algorithm.

The partial derivative $\partial J_{\rm e}(\mathcal{Y})/\partial\mathbf{w}^*_{a,m}$ can be expressed as

$$\frac{J_{\rm e}(\mathcal{Y})}{\partial\mathbf{w}^*_{a,m}} = \frac{1}{K}\sum_{k=0}^{K-1}\left[\frac{\partial J_{\rm e}(Y(k))}{\partial|Y_m(k)|}\frac{\partial|Y_m(k)|}{\partial\mathbf{w}^*_{a,m}} + \frac{\partial J_{\rm e}(Y(k))}{\partial\hat{\sigma}_m(k)}\frac{\partial\hat{\sigma}_m(k)}{\partial\mathbf{w}^*_{a,m}}\right],$$

where $\mathcal{Y} = \{Y(k)\}_{k=0}^{K-1}$ is the set of data used to adapt the beamforming weights. The first term in the sum can be expressed as

$$\begin{aligned}&\frac{\partial J_{\rm e}(Y(k))}{\partial|Y_m(k)|}\cdot\frac{\partial|Y_m(k)|}{\partial\mathbf{w}^*_{a,m}}\\&\quad= \frac{1}{K}\left\{\frac{f\,|Y(k)|^{f-2}}{2\,[B_{\rm c}(p)\,\hat{\sigma}_m(k)]^f} - \frac{1}{\hat{\sigma}^2_m(k)}\right\}\mathbf{B}_m^H\,\mathbf{X}_m(k)\,Y_m^*(k).\end{aligned}$$

Following the definition (11), the first part of the other term can be expressed as

$$\frac{\partial J_{\rm e}(Y(k))}{\partial\hat{\sigma}_m(k)} = \frac{1}{K}\left[\frac{\partial\log p_{\rm gg}(Y(k))}{\partial\hat{\sigma}_m(k)} - \frac{\partial\log p_{\rm Gauss}(Y(k))}{\partial\hat{\sigma}_m(k)}\right],$$

where the details of the derivation are provided in [12]. Based on (3), we can write

$$\begin{aligned}&\frac{\partial\log p_{\rm gg}(Y(k);f,\hat{\sigma}_m(k))}{\partial\hat{\sigma}_m(k)}\cdot\frac{\partial\hat{\sigma}_m(k)}{\partial\mathbf{w}^*_{a,m}} =\\&\quad\frac{1}{2}\cdot\left[\left(\frac{f\,|Y(k)|^f}{B_{\rm c}^f(f)\,\hat{\sigma}_m^{f+1}(k)} - \frac{2}{\hat{\sigma}_m(k)}\right)\cdot\hat{\sigma}_m(k)\cdot\frac{\partial\bar{\eta}_m}{\partial\mathbf{w}^*_{a,m}}\right],\end{aligned}$$

where $\bar{\eta}_m$ denotes the mth component of $\bar{\boldsymbol{\eta}}$. Based on (10), we can then write

$$\begin{aligned}\frac{\partial\bar{\eta}_m}{\partial\mathbf{w}^*_{a,m}} &= \frac{\mathbf{t}'_m\,\mathbf{t}_m}{K}\sum_{k'=0}^{K-1}\frac{1}{|Y_m(k')|^2}\cdot\frac{\partial|Y_m(k')|^2}{\partial\mathbf{w}^*_{a,m}}\\&= \frac{-\mathbf{t}'_m\,\mathbf{t}_m}{K}\sum_{k'=0}^{K-1}\frac{1}{|Y_m(k')|^2}\cdot\mathbf{B}_m^H\,\mathbf{X}_m(k')\,Y_m^*(k'),\end{aligned} \tag{12}$$

where $\mathbf{t}'_m$ is the mth *row* of $\mathbf{T}_{\nu}^{-1}$ and $\mathbf{t}_m$ is the mth *column* of $\mathbf{T}_{\nu}$. In writing (12), we account only for the effect of $\mathbf{w}^*_{{\rm a},m}$ on $\bar{\eta}_m$, and ignore its effect on any other $\bar{\eta}_i$ for $i \neq m$. Thus we are acting on the assumption that $\mathbf{T}_{\nu}^{-1}\mathbf{T}_{\nu}$ is *diagonally dominated*. Let $\hat{\sigma}^2_m(k)$ denote the mth diagonal component of $\boldsymbol{\Sigma}_{\mathbf{Y}}(k)$. It then follows that

$$\frac{\partial\hat{\sigma}_m(k)}{\partial\mathbf{w}^*_{a,m}} = \frac{1}{2}\cdot\exp\left(\frac{1}{2}\eta_m(k)\right)\cdot\frac{\partial\bar{\eta}_m}{\partial\mathbf{w}^*_{a,m}} = \frac{\hat{\sigma}_m(k)}{2}\cdot\frac{\partial\bar{\eta}_m}{\partial\mathbf{w}^*_{a,m}}.$$

5. EXPERIMENTS

We performed far-field ASR experiments on the MC-WSJ-AV; see [13] for a description of the data collection apparatus. In the single speaker stationary scenario of the MC-WSJ-AV, a speaker was asked to sit or stand in front of a presentation screen and read sentences from different positions. The far-field speech data was recorded with two circular, eight-channel microphone arrays in a reverberant room. In addition to the reverberation, some recordings include significant amounts of background noise. Our test data set for the experiments contains 10 speakers recorded with the first array where each speaker reads approximately 40 sentences taken from the 5,000 word vocabulary WSJ task. This provided a total of 352 utterances which correspond to approximately 43.9 minutes of speech. There are a total of 11,598 word tokens in the reference transcriptions. Prior to beamforming, we first estimated the speaker's position with an automatic source tracking system [14, 15]. Based on the average speaker position estimated for each utterance, utterance-dependent active weight vectors $\mathbf{w}_a$ were estimated for the source.

Iterations of the conjugate gradients algorithm were run on the entire utterance until convergence was achieved. We did four decoding passes on the waveforms obtained with the beamforming algorithms described above. Each pass of decoding used a different acoustic model or speaker adaptation scheme. Speaker adaptation parameters were estimated using the word lattices generated during the prior pass. The details of the speech recognition engine are presented in [15].

Beamforming	Pass (%WER)			
Algorithm	1	2	3	4
D&S BF	80.1	39.9	21.5	17.8
MNB, global Variance	75.3	34.8	18.2	**14.6**
HMM-MNB	74.9	32.7	16.9	**13.6**
HMM-MNB, oracle alignments	75.0	33.7	17.2	14.1
Single distant microphone	87.0	57.1	32.8	28.0
Close talking microphone	52.9	21.5	9.8	6.7

Table 1. WERs after every decoding pass.

Table 5 shows the word error rates (WERs) for every beamforming algorithm. As references, WERs in recognition experiments on speech data recorded with the single distant microphone and with the close-talking microphone (CTM) are also given in Table 5. We observe that the HMM-MNB performs better than the global variance baseline on all passes, with a 1% absolute gain on the final pass. The results of the oracle setup are also better than the baseline, but worse than the non-oracle case. This may seem surprising at first since the real transcriptions could be expected to lead to more accurate speech modelling. We suspect that the superior performance of the non-oracle case is due to the fact that the incorrect transcription stems from a better match between those models and the data, which in turn leads to a more accurate reconstruction of the PSD.

6. CONCLUSIONS AND FUTURE WORK

We have presented an extension to the maximum negentropy beamformer that aims to account for the non-stationarity of speech. This beamforming algorithm uses an auxiliary HMM to model the short-term variation of speech during beamforming. In a set of far-field ASR experiments on data from the Multi-Channel Wall Street Journal Audio-Visual Corpus, we were able to reduce the word error rate from 14.6% to 13.6% by accounting for this non-stationarity. Future work will consider the use of phone and state-dependent shape factors for the GG pdfs considered here.

7. REFERENCES

[1] M. L. Seltzer, B. Raj, and R. M. Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *IEEE Trans. Speech Audio Proc.*, vol. 12, no. 5, pp. 489–498, 2004.

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

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

[4] Kenichi Kumatani, John McDonough, Dietrich Klakow, Philip N. Garner, and Weifeng Li, "Adaptive beamforming with a maximum negentropy criterion," in *Proc. of HSCMA*, submitted February 2008.

[5] Kenichi Kumatani, John McDonough, Dietrich Klakow, Philip N. Garner, and Weifeng Li, "Adaptive beamforming with a maximum negentropy criterion," *IEEE Trans. on ASLP*, submitted February 2008.

[6] 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.

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

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

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

[10] Matthias Wölfel and John McDonough, *Distant Speech Recognition*, Wiley, New York, 2008.

[11] J.W. McDonough, *Speaker compensation with all–pass transforms*, Doctor thesis, Johns Hopkins University, Baltimore, Maryland, USA, 2000.

[12] Barbara Rauch, Kenichi Kumatani, and John McDonough, "Hidden Markov model beamforming with a maximum negentropy optimization criterion," Tech. Rep., Spoken Language Systems, Saarland University, Saarbrücken, Germany, March 2008.

[13] 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. of ASRU*, 2005, pp. 357–362.

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

[15] John McDonough, Kenichi Kumatani, Tobias Gehrig, Emilian Stoimenov, Uwe Mayer, Stefan Schacht, Matthias Woelfel, and Dietrich Klakow, "To separate speech! a system for recognizing simultaneous speech," in *Proc. MLMI*, 2007.