MAXIMUM NEGENTROPY BEAMFORMING WITH SUPERDIRECTIVITY

Kenichi Kumatani[1], *Liang Lu*[2], *John McDonough*[1], *Arnab Ghoshal*[1] *and Dietrich Klakow*[1]

[1]Spoken Language Systems at Saarland University in Saarbrücken, Germany
[2]The Centre for Speech Technology Research at University of Edinburgh in Edinburgh, United Kingdom
web : http://distant-automatic-speech-recognition.org

ABSTRACT

This paper presents new superdirective beamforming algorithms based on the maximum negentropy (MN) criterion for distant automatic speech recognition. The MN beamformer is configured in the *generalized sidelobe canceler* structure, and uses the weights derived from a delay-and-sum beamformer as the *quiescent weight vector*. While satisfying the distortionless constraint in the look direction, it adjusts the *active weight vector* to make the output maximally super-Gaussian.

The current paper proposes to use the weights of a superdirective beamformer as the quiescent vector, which results in improved directivity and noise suppression at lower frequencies. We demonstrate the effectiveness of our approach through far-field speech recognition experiments on the *Multi-Channel Wall Street Journal Audio Visual Corpus* (MC-WSJ-AV). The technique proposed in the current paper reduces the word error rate (WER) by 56% relative to a single distant microphone baseline, which is a 14% reduction in WER over the original MN beamformer formulation.

1. INTRODUCTION

Microphone array processing techniques for hands-free speech recognition have the potential to relieve users from the necessity of donning close talking microphones (CTMs) before dictating or otherwise interacting with automatic speech recognition (ASR) systems [1, 2].

Adaptive beamforming is a promising technique for far-field speech recognition. A conventional beamformer in *generalized sidelobe canceller* (GSC) configuration is structured such that the direct signal from a desired direction is undistorted [2, §6.7.3]. Typical GSC beamformers consist of three blocks, a *quiescent vector*, *blocking matrix* and *active weight vector*. The quiescent vector is calculated to provide unity gain for the direction of interest. The blocking matrix is usually constructed in order to keep a distortionless constraint for the signal filtered with the quiescent vector. Subject to the constraint, the total output power of the beamformer is minimized through the adjustment of an active weight vector, which effectively places a null on any source of interference, but can also lead to undesirable *signal cancellation* [3]. To avoid the latter, many algorithms have been developed. Those approaches could be classified into the following :

1. updating the active weight vector only when noise signals are dominant [4],
2. constraining the update formula for the active weight vector [5],
3. blocking the leakage of desired signal components into the sidelobe canceller by designing the blocking matrix [5, 6], and
4. using acoustic transfer functions from a desired source to microphones instead of just compensating time delays [4, 6].

Those algorithms attempt to minimize the almost same criterion based on the the second-order statistics (SOS), the total output power while keeping the distortionless constraint.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 213850 and the Cluster of Excellence on Multimodal Computing and Interaction.

We know from the field of independent component analysis (ICA) that nearly all information bearing signals, like subband samples of speech, are *non*-Gaussian [7]. On the other hand, noisy or reverberant speech consist of a sum of several signals, and as such tend to have a distribution that is closer to Gaussian. This follows from the *central limit theorem*, and can be empirically verified [8]. Hence, by making the distribution of the beamformer's outputs as much non-Gaussian as possible, we can remove the effects of noise and reverberation.

In [8], we proposed a novel beamforming algorithm which adjusted the active weight vectors so as to make the beamformer's output maximally non-Gaussian. As a measure for the degree of non-Gaussianity we use negentropy, which is the difference between the entropy of the output signal calculated under a Gaussian assumption and that calculated under a non-Gaussian assumption. In other words, negentropy is a measure for the amount by which the distribution of the beamformer's output deviates from a Gaussian with the same mean and variance. We also showed in [8] that such a beamformer can reduce noise and reverberation without suffering from the signal cancellation problem.

The MN beamformer proposed in [8] used the weights of a delay-and-sum beamformer, which compensates time delays of arrival of a desired speech signal to the microphone array, as the quiescent vector. However, due to the limited aperture of the microphone array, such a delay-and-sum beamforming method cannot suppress interference signals at low frequencies. Since the output of the quiescent vector influences the negentropy of the beamformer's output, presence of noise in that output degrades the ability of the beamformer to suppress noise or reverberation by estimating the active weight vector based on the maximum negentropy criterion. A superdirective beamformer alleviates this problem by having better directivity at lower frequencies.

The balance of this paper is organized as follows. Section 2 reviews the super-Gaussian distribution and shows the fact that the actual speech distribution is not Gaussian but super-Gaussian, which is the main motivation for using the maximum negentropy criterion. In Section 3 and Section 4, we review the definitions of the entropy and negentropy, respectively. In Section 5, we describe the super-directive beamformer. Section 6 describes the new maximum negentropy beamformer in the GSC configuration. In Section 7, we describe the results of far-field automatic speech recognition experiments. Finally, in Section 8, we present our conclusions and plans for future work.

2. MODELING SUBBAND SAMPLES OF SPEECH WITH SUPER-GAUSSIAN PROBABILITY DENSITY FUNCTIONS

In this section we provide empirical evidence that the probability density function (pdf) of speech is super-Gaussian. We use a generalized Gaussian pdf to model the distribution of the subband speech samples.

2.1 Generalized Gaussian pdf

The generalized Gaussian (GG) pdf is well-known and finds frequent application in the blind source separation (BSS) and ICA fields. Moreover, it subsumes the Gaussian and Laplace pdfs as special cases. The GG pdf with zero mean for a real-valued r.v. y

can be expressed as

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

where

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

In (2), $\Gamma(.)$ is the gamma function and p is the shape parameter, which controls how fast the tail of the pdf decays. Note that the GG with $p = 1$ corresponds to the Laplace pdf, and that setting $p = 2$ yields the Gaussian pdf, whereas in the case of $p \to +\infty$ the GG pdf converges to a uniform distribution.

The maximum likelihood (ML) estimation is a straightforward method for estimating parameters of pdfs. For a set $\mathscr{Y} = \{y_0, y_1, \ldots, y_{N-1}\}$ of N real-valued training samples, the ML solution of the scale parameter can be expressed in [8, 9] as

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

Notice that the ML solution of the scale parameter $\hat{\sigma}$ is different from the variance in the case of $p \neq 2$.

Due to the presence of the special functions, it is impossible to solve the log-likelihood function for p explicitly. Varanasi [9] showed, however, that there is a unique root given the scale parameter. Hence, we use the linear search in order to find the solution of the shape parameter. During beamforming, the shape parameter is fixed and negentropy described in Section 4 is maximized based on (3).

2.2 Super-Gaussian Characteristics of Clean Speech

The Gaussian and four super-Gaussian univariate pdfs considered are plotted in Fig. 1, where the parameters of the generalized Gaussian (GG) pdf are calculated from samples of actual speech subbands. From the figure, it is clear that the Laplace, K_0, Γ and GG pdfs exhibit the "spikey" and "heavy-tailed" characteristics that are typical of super-Gaussian pdfs. This implies that they have a sharp concentration of probability mass at the mean, relatively little probability mass as compared with the Gaussian at intermediate values of the argument, and a relatively large amount of probability mass in the tail; i.e., far from the mean.

Fig. 1 also shows the histogram of the real parts of subband components at f_s = 800 Hz. To generate these histograms, we used 43.9 minutes of clean speech recorded with the CTM in the development set of the Speech Separation Challenge, Part 2 (SSC2) [1].

Fig. 2 shows histograms of real parts of subband components calculated from clean speech and noise corrupted speech. It is clear from this figure that the pdf of the noise-corrupted signal, which is in fact the sum of the speech and noise signals, is closer to Gaussian than that of clean speech. Fig. 3 shows histograms of clean speech and reverberated speech in the subband domain. In order to produce reverberated speech, a clean speech signal was convolved with an impulse response measured in a room; see Lincoln *et al.* [1] for the configuration of the room. We can observe from Fig. 3 that the pdf of reverberated speech is also closer to Gaussian than the original clean speech.

These facts would indeed support the hypothesis that seeking an enhanced speech signal that is maximally non-Gaussian is an effective way to suppress the distorting effects of noise and reverberation.

3. ENTROPY

Entropy is the basic measure of information in *information theory* [7]. The entropy for a continuous r.v. Y, which is often called the differential entropy, is defined as

$$H(Y) \triangleq -\int p_Y(v)\log p_Y(v)dv = -\mathscr{E}\{\log p_Y(v)\}, \qquad (4)$$

where $p_Y(.)$ is the pdf of Y. The entropy of a r.v. is a measure of the uncertainty associated with the r.v. Accordingly, large entropy indicates that the variables contain unstructured information.

The entropy for a Gaussian r.v. Y_{gauss} can be expressed as

$$H(Y_{\mathrm{gauss}}) = \log\left|\sigma_Y^2\right| + (1+\log\pi). \qquad (5)$$

where σ_Y^2 is the variance of the r.v.s. A Gaussian r.v. has the largest entropy among all r.v.s of equal variance [7]. 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*.

The differential entropy of the GG pdf for the real-valued r.v. y is obtained as

$$\begin{aligned} H_{\mathrm{GG}}(y) \quad &= \quad -\int_{-\infty}^{+\infty} p_{\mathrm{gg}}(\xi)\log p_{\mathrm{gg}}(\xi)d\xi \\ &= \quad \frac{1}{p} + \log\left[2\Gamma(1+1/p)A(p)\hat{\sigma}\right]. \end{aligned} \qquad (6)$$

4. NEGENTROPY

Negentropy is frequently used in order to measure nongaussianity in the field of ICA. Negentropy is the distance between the entropy of Gaussian and non-Gaussian r.v.s. In this work, we use the GG pdf for the real-valued r.v. and calculate negentropy as

$$J(Y) = H(Y_{\mathrm{gauss}}) - H_{\mathrm{GG}}(|Y|). \qquad (7)$$

where Y_{gauss} is a Gaussian variable which has the same variance σ_Y^2 as Y. Note that negentropy is non-negative, and it is minimum if and only if Y has a Gaussian distribution.

Kurtosis is also used for measuring non-Gaussianity. The kurtosis criterion does not require any pdf assumption. Due to its simplicity, it is widely used as a measure of non-Gaussianity. However, the value of kurtosis might be greatly influenced by a few samples with a low observation probability. Hyvärinen and Oja [7] noted that negentropy was generally more robust in the presence of outliers than kurtosis. We also applied the maximum kurtosis criterion to beamforming and confirmed that maximum negentropy beamforming is more robust for the outliers [10].

5. SUPER-DIRECTIVE BEAMFORMING

To describe super-directive beamforming, we start with the explanation of the minimum variance distortionless response (MVDR) beamforming.

The MVDR beamforming algorithm determines the optimum weight vector that minimizes the beamformer's output at each frequency bin m:

$$\mathbf{w}^H(m)\mathbf{\Sigma_N}(m)\mathbf{w}(m), \qquad (8)$$

subject to the distortionless constraint for the desired look direction

$$\mathbf{w}^H(m)\mathbf{d}(m) = 1, \qquad (9)$$

where $\mathbf{d}(m)$ is the beam-steering vector and $\mathbf{\Sigma_N}$ is the spatial spectral matrix of noise. The well-known solution is called the minimum variance distortionless response (MVDR) beamformer [2, §13]. The weight vector of the MVDR beamformer can be expressed as

$$\mathbf{w}_{\mathrm{MVDR}}(m) = \frac{\mathbf{\Sigma_N}^{-1}(m)\mathbf{d}(m)}{\mathbf{d}^H(m)\mathbf{\Sigma_N}^{-1}(m)\mathbf{d}(m)}. \qquad (10)$$

The MVDR beamfomers would attempt to null out any interfering signal, but are prone to the signal cancellation problem [3] whenever there is an interfering signal that is correlated with the desired signal. In realistic environments, interference signals are highly correlated with a target signal since the target signal is reflected from hard surfaces such as walls and tables. Therefore, the

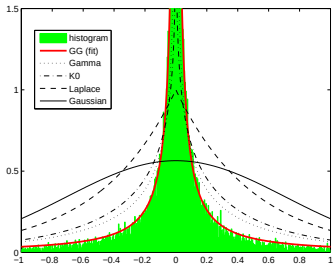

Figure 1: Histogram of real parts of subband components and the likelihood of pdfs.

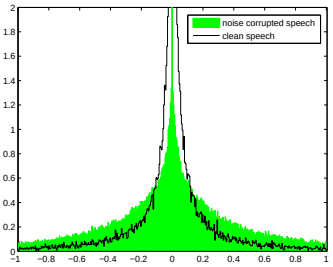

Figure 2: Histograms of clean speech and noise corrupted speech in the subband domain.

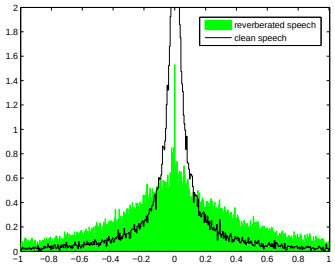

Figure 3: Histograms of clean speech and reverberated speech in the subband domain.

adaptation of the weight vector is usually halted whenever the desired source is active.

Instead of using noise observations, we can model noise fields. Assuming that the noise field is spherical isotropic or diffuse, the spatial spectral matrix of noise $\mathbf{\Sigma_N}$ can be replaced with the coherence matrix $\mathbf{\Gamma_N}$. The (i,j)-th component of the coherence matrix can be expressed as

$$\mathbf{\Gamma}_{\mathbf{N}ij}(m) = \mathrm{sinc}\left(\frac{2\pi d_{ij} f_s}{c}\frac{m}{M}\right), \tag{11}$$

where d_{ij} is the distance between the i-th and j-th elements of the array, f_s is the sampling frequency, c is the sound speed and M is the number of subbands. Now, we can express the weights of the super-directive beamformer as

$$\mathbf{w}_{\mathrm{SD}}(m) = \frac{\mathbf{\Gamma}_{\mathbf{N}}^{-1}(m)\mathbf{d}(m)}{\mathbf{d}^H(m)\mathbf{\Gamma}_{\mathbf{N}}^{-1}(m)\mathbf{d}(m)}. \tag{12}$$

The beamformer which optimizes the directivity factor with the ratio of the wavelength to the distance between the sensors is termed the *super-directive beamformer*.

Figure 4 shows the beam patterns of the super-directive and delay-and-sum beamformers as a function of the azimuth and frequency, where the look direction is 0 radian. It is clear from Figure 4 that the delay-and-sum beamformer is unable to suppress interference signals at low frequencies while the super-directive beamformer can form a sharp beam for the look direction.

Additional weight is typically added to the main diagonal of $\mathbf{\Gamma_N}$ in order to avoid excessively large sidelobes in the beam pattern and the attendant non-robustness [2, §13]. The same effect can be also obtained by dividing the non-diagonal elements instead of adding diagonal elements [11, §2]. In this work, we divide non-diagonal elements by 1.01.

Notice that the coherence matrix of the super-directive beamformer is decided by the geometry of the microphone array only and does not suffer the signal cancellation problem.

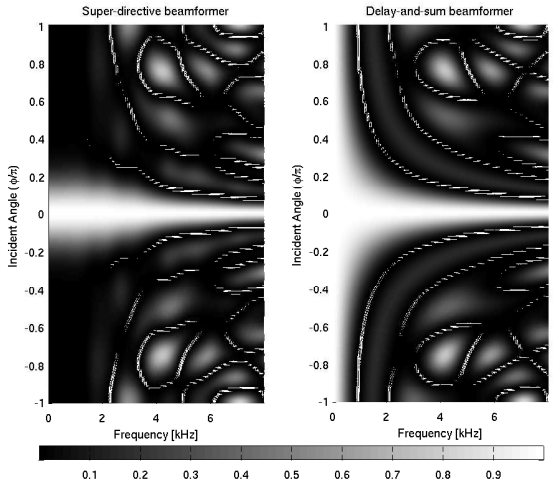

Figure 4: Beam patterns of the super-directive and delay-and-sum beamformers.

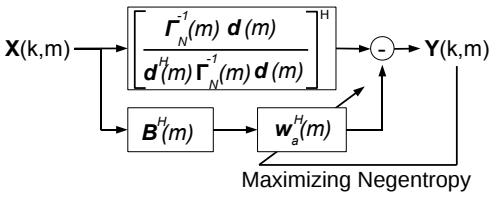

Figure 5: Schematic of the proposed GSC beamformer.

6. SUPER-DIRECTIVE BEAMFORMING BASED ON THE MAXIMUM NEGENTROPY CRITERION

Consider a subband beamformer in the GSC configuration [2, §13.7.3], as shown in Fig. 5. The output of our beamformer for a given subband m can be expressed as

$$Y(k,m) = \left(\mathbf{w}_{\mathrm{SD}}(m) - \mathbf{B}(m)\mathbf{w}_{\mathrm{a}}(m)\right)^H \mathbf{X}(k,m), \tag{13}$$

where $\mathbf{w}_{\mathrm{SD}}(m)$ is the *quiescent weight vector* for a source, $\mathbf{B}(m)$ is the *blocking matrix*, $\mathbf{w}_{\mathrm{a}}(m)$ is the *active weight vector*, and $\mathbf{X}(k,m)$ is the input subband *snapshot vector* at frame k.

In contrast to conventional algorithms, the super-directive beamformer's weights expressed in (12) are used as the *quiescent weight vector*. In this work, we consider two kinds of orthogonal conditions for blocking matrices: $\mathbf{B}^H(m)\cdot\mathbf{d}(m) = \mathbf{0}$ (type1) and $\mathbf{B}^H(m)\cdot\mathbf{w}_{\mathrm{SD}}(m) = \mathbf{0}$ (type2).

Both of them can keep the distortionless constraint for the look direction. However, the areas which can be controlled by the sidelobe canceler are different. This orthogonality implies that the distortionless constraint will be satisfied for any choice of $\mathbf{w}_{\mathrm{a}}$. The blocking matrix can be calculated with an orthogonalization technique such as the modified Gram-Schmidt, QR decomposition or singular value decomposition [12]. In this work, we used the modified Gram-Schmidt orthogonalization technique. While the active weight vector $\mathbf{w}_{\mathrm{a}}$ is typically chosen to minimize the variance of the beamformer's outputs which leads to the undesired signal cancellation, here we will develop an optimization procedure to find that $\mathbf{w}_{\mathrm{a}}$ which maximizes the negentropy $J(Y)$ described in Section 4.

In conventional GSC beamforming, the *regularization term* is often applied in order to penalize large active weight vectors, and

thereby improve robustness by inhibiting the formation of excessively large sidelobes [2, §13.3.8]. Such a regularization term can be applied in the present instance by defining the modified optimization criterion

$$\mathscr{J}(Y;\alpha) = J(Y) - \alpha\|\mathbf{w}_{\mathrm{a}}\|^2 \tag{14}$$

for some real $\alpha > 0$. In our previous work [8], we did not observe a significant effect of parameter α. Accordingly, we set $\alpha = 0.01$ in this work which provided the best result in [8].

For the experiments described in Section 7, subband analysis and synthesis were performed with a uniform DFT filter bank based on the modulation of a single prototype impulse response [13], which was designed to minimize each aliasing term individually. Beamforming in the subband domain has the considerable advantage that the active sensor weights can be optimized for each subband independently, which provides a tremendous computational saving with respect to a time-domain filter-and-sum beamformer with filters of the same length on the output of each sensor.

6.1 Estimation of Active Weights under the Generalized Gaussian pdf

Due to the absence of the close-form solution, we resort to the numerical optimization algorithms such as the conjugate gradients algorithm in order to obtain the active weight vectors. In this section, we omit the frequency index *m* for the sake of simplicity.

In order to calculate the negentropy, we first need the variance of the beamformer outputs $Y(k)$. Substituting (13) into the definition $\sigma_Y^2 = \mathscr{E}\{YY^*\}$ of variance, we find

$$\sigma_Y^2 = (\mathbf{w}_{\mathrm{SD}} - \mathbf{B}\mathbf{w}_{\mathrm{a}})^H \boldsymbol{\Sigma}_{\mathbf{X}} (\mathbf{w}_{\mathrm{SD}} - \mathbf{B}\mathbf{w}_{\mathrm{a}}), \tag{15}$$

where $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathscr{E}\{\mathbf{X}\mathbf{X}^H\}$ is the covariance matrix of the input snapshot vectors.

In order to apply the conjugate gradients algorithm, we must derive an expression for the gradient. By substituting (5) and (6) into (14) and taking the partial derivative on both sides while holding the shape parameter fixed, we obtain

$$\frac{\partial \mathscr{J}(Y;\alpha)}{\partial \mathbf{w}_{\mathrm{a}}{}^*} = -\frac{1}{K}\sum_{k=0}^{K-1}\left\{\frac{1}{\sigma_Y^2} - \frac{p|Y(k)|^{p-2}}{2\,|A(p)\hat{\sigma}|^p}\right\}\mathbf{B}^H\mathbf{X}(k)Y^*(k) - \alpha\mathbf{w}_{\mathrm{a}}. \tag{16}$$

where $\hat{\sigma}$ is calculated with (3).

Based on (15) through (16), a numerical algorithm for optimizing the active weight vector can be implemented.

7. EXPERIMENTS

We performed far-field automatic speech recognition (ASR) experiments on the *Multi-Channel Wall Street Journal Audio Visual Corpus* (MC-WSJ-AV) from the *Augmented Multi-party Interaction* (AMI); see Lincoln et al. [1] for the detail of the data collection apparatus. The room size is 650 cm × 490 cm × 325 cm and the reverberation time T_{60} was approximately 380 millisecond. In addition to reverberation, some recordings include significant amounts of background noise such as computer fan and air conditioner noise. The far-field speech data was recorded with two circular, equi-spaced eight-channel microphone arrays with diameter of 20cm. Additionally, the close talking headset microphone (CTM) is used for each speaker. The sampling rate of the recordings was 16 kHz. As the data was recorded with real speakers in a realistic acoustic environment and not artificially convolved with measured room impulse responses, the positions of the speakers' heads as well as the speaking volume vary even though the speakers are largely stationary. Indeed, it is exactly this behavior of real speakers that makes working with data from corpora such as MC-WSJ-AV so much more challenging than working with data that was played through a loud speaker into a room, not to mention data that was *artificially convolved*. In the *single speaker stationary* scenario of the MC-WSJ-AV, a speaker was asked to read sentences from six positions, four seated around the table, one standing at the white board and one standing at presentation screen.

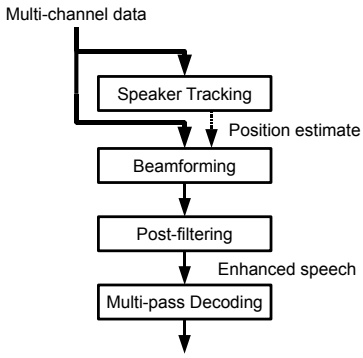

Figure 6: A block chart of the distant speech recognition system.

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

Figure 6 illustrates our distant speech recognition system for experiments. Prior to beamforming, we first estimated the speaker's position with the *Orion* source tracking system [14]. Based on the average speaker position estimated for each utterance, utterance-dependent active weight vectors $\mathbf{w}_a$ were estimated for a source. The active weight vectors for each subband were initialized to zero for estimation. Iterations of the conjugate gradients algorithm were run on the entire utterance until convergence was achieved. Zelinski post-filtering [15] was performed after beamforming. The parameters of the GG pdf were trained with 43.9 minutes of speech data recorded with the CTM in the SSC development set. The training data set for the GG pdf contains recordings of 5 speakers.

We performed four decoding passes on the waveforms obtained with each of the beamforming algorithms described in prior sections. The details of our ASR system used in the experiments are written in [8]. Each pass of decoding used a different acoustic model or speaker adaptation scheme. For all passes save the first unadapted pass, speaker adaptation parameters were estimated using the word lattices generated during the prior pass, as in [16]. A description of the four decoding passes follows:

1. Decode with the unadapted, conventional ML acoustic model and bigram language model (LM).
2. Estimate vocal tract length normalization (VTLN) [17] parameters and constrained maximum likelihood linear regression parameters (CMLLR) [18] for each speaker, then redecode with the conventional ML acoustic model and bigram LM.
3. Estimate VTLN, CMLLR, and maximum likelihood linear regression (MLLR) [19] parameters for each speaker, then redecode with the conventional model and bigram LM.
4. Estimate VTLN, CMLLR, MLLR parameters for each speaker, then redecode with the ML-SAT model and bigram LM.

Table 1 shows the word error rates (WERs) for every beamforming algorithm. As references, WERs in recognition experiments on speech data recorded with the single distant microphone (SDM) and CTM are described in Table 1.

It is clear from Table 1 that the best recognition performance, WER 12.1%, is obtained by super-directive beamforming based on the maximum negentropy criterion with the orthogonal condition $\mathbf{B}^H(m)\cdot\mathbf{w}_{\mathrm{SD}}(m) = \mathbf{0}$ (SD-MN BF (type2)). It is also clear from Table 1 that the super-directive beamforming algorithm with the other orthogonal condition $\mathbf{B}^H(m)\cdot\mathbf{d}(m) = \mathbf{0}$ (SD-MN BF (type1)) provides the second best recognition performance, WER 12.4%

It can be seen from Table 1 that conventional maximum beam-

Table 1: Word error rates for each beamforming algorithm after every decoding pass.

Beamforming Algorithm	Pass (%WER)			
	1	2	3	4
D&S BF	79.0	38.1	20.2	16.5
MVDR BF	78.6	35.4	18.8	14.8
SD BF	71.4	31.9	16.6	14.1
GEV BF	78.7	35.5	18.6	14.5
Conventional MN BF	75.1	32.7	16.5	13.2
SD-MN BF (type1)	74.1	30.8	15.2	12.4
SD-MN BF (type2)	74.9	32.1	15.4	12.1
SDM	87.0	57.1	32.8	28.0
CTM	52.9	21.5	9.8	6.7

forming algorithm (Conventional MN BF) can provide the better recognition performance than the other traditional beamforming methods, the delay-and-sum beamformer (D&S BF), superdirective beamformer (SD-BF) and the minimum variance distortionless response (MVDR) beamformer. Notice MVDR beamforming algorithms require speech activity detection in order to avoid the signal cancellation. For the adaptation of the MVDR beamformer, we used the first 0.1 and last 0.1 seconds in each utterance data which contain only background noise. Again, in contrast to conventional beamforming methods, our algorithm does not need to detect the start and end points of target speech since the proposed method can suppress noise and reverberation without the signal cancellation problem.

Table 1 also shows the recognition results obtained with the generalized eigenvector beamformer (GEV BF) proposed by E. Warsitz et al. [6]. It achieved slightly better recognition performance than the MVDR beamformer. In this task, the transfer function from the sound source to the microphone array changes in time due to movements of the speaker's head. Moreover, it is difficult to determine whether or not the signal observed at any given time contains both speech and noise components in each frequency bin, which is required to estimate the transfer function. Due to these difficulties, the performance of the GEV beamformer is limited in realistic environments. It is worth noting that the best result of 12.1% in Table 1 is significantly less than half the word error rate reported elsewhere in the literature on this far-field ASR task [1].

8. CONCLUSIONS AND FUTURE WORK

In this work, we have proposed novel super-beamforming algorithms based on maximizing negentropy. The proposed methods do not exhibit the signal cancellation problems typically seen in conventional adaptive beamformers. We also evaluated the beamforming algorithms through a set of far-field automatic speech recognition experiments on the data captured in realistic acoustic environments and spoken by real speakers. In these experiments, the superdirective beamformer with the maximum negentropy criterion provided the best ASR performance.

We plan to develop an on–line version of the beamforming algorithm presented here. This on–line algorithm will be capable of adjusting the active weight vectors $\mathbf{w}_a$ with each new snapshot in order to track changes of speaker position and movements of the speaker's head during an utterance.

REFERENCES

[1] M. Lincoln, I. McCowan, I. Vepa, and H. K. Maganti, "The multi-channel Wall Street Journal audio visual corpus (mc-wsj-av): Specification and initial experiments," in *Proc. IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2005, pp. 357–362.

[2] M. Wölfel and J. McDonough, *Distant Speech Recognition*. London: Wiley, 2009.

[3] B. Widrow, K. M. Duvall, R. P. Gooch, and W. C. Newman, "Signal cancellation phenomena in adaptive antennas: Causes and cures," *IEEE Transactions on Antennas and Propagation*, vol. AP-30, pp. 469–478, 1982.

[4] I. Cohen, S. Gannot, and B. Berdugo, "An integrated real-time beamforming and postfiltering system for nonstationary noise environments," *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 1064–1073, 2003.

[5] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Transactions on Signal Processing*, vol. 47, pp. 2677–2684, 1999.

[6] E. Warsitz, A. Krueger, and R. Häb-Umbach, "Speech enhancement with a new generalized eigenvector blocking matrix for application in a generalized sidelobe canceller," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, U.S.A, 2008.

[7] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, 2000.

[8] K. Kumatani, J. McDonough, B. Rauch, D. Klakow, P. N. Garner, and W. Li, "Beamforming with a maximum negentropy criterion," *IEEE Trans. Audio, Speech and Language Processing*, vol. 17, pp. 994–1008, 2009.

[9] M. K. Varanasi, "Parameter estimation for the generalized gaussian noise model," Ph.D. dissertation, Rice University, 1987.

[10] K. Kumatani, J. McDonough, B. Rauch, P. N. Garner, W. Li, and J. Dines, "Maximum kurtosis beamforming with the generalized sidelobe canceller," in *Proc. Interspeech*, Brisbane, Australia, 2008.

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

[12] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore: The Johns Hopkins University Press, 1996.

[13] K. Kumatani, J. McDonough, S. Schacht, D. Klakow, P. N. Garner, and W. Li, "Filter bank design based on minimization of individual aliasing terms for minimum mutual information subband adaptive beamforming," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, Nevada, U.S.A, 2008.

[14] T. Gehrig, U. Klee, J. McDonough, S. Ikbal, M. Wölfel, and C. Fügen, "Tracking and beamforming for multiple simultaneous speakers with probabilistic data association filters," in *Proc. Interspeech*, 2006, pp. 2594–2597.

[15] C. Marro, Y. Mahieux, and K. U. 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.

[16] L. Uebel and P. Woodland, "Improvements in linear transform based speaker adaptation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001.

[17] L. Welling, H. Ney, and S. Kanthak, "Speaker adaptive modeling by vocal tract normalization," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 415–426, 2002.

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

[19] 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.