Maximum kurtosis beamforming with a subspace filter for distant speech recognition

Kenichi Kumatani [#1], John McDonough [*2], Bhiksha Raj [*3]

[#] *Disney Research, Pittsburgh*
4720 Forbes Ave, Lower Level, Suite 110, Pittsburgh, PA 15231, USA
[1] k_kumatan@ieee.org

[*] *Carnegie Mellon University*
5000 Forbes Avenue, GHC, Pittsburgh, PA 15213, USA
[2,3] {johnmcd,bhiksha}@cs.cmu.edu

***Abstract*—This paper presents a new beamforming method for distant speech recognition (DSR). The *dominant mode* subspace is considered in order to efficiently estimate the *active weight vectors* for maximum kurtosis (MK) beamforming with the *generalized sidelobe canceler* (GSC). We demonstrated in [1], [2], [3] that the beamforming method based on the maximum kurtosis criterion can remove reverberant and noise effects without signal cancellation encountered in the conventional beamforming algorithms. The MK beamforming algorithm, however, required a relatively large amount of data for reliably estimating the active weight vector because it relies on a numerical optimization algorithm. In order to achieve efficient estimation, we propose to cascade the subspace (eigenspace) filter [4, §6.8] with the active weight vector. The subspace filter can decompose the output of the *blocking matrix* into directional signals and ambient noise components. Then, the ambient noise components are averaged and would be subtracted from the beamformer's output, which leads to reliable estimation as well as significant computational reduction. We show the effectiveness of our method through a set of distant speech recognition experiments on real microphone array data captured in the real environment. Our new beamforming algorithm provided the best recognition performance among conventional beamforming techniques, a word error rate (WER) of 5.3 %, which is comparable to the WER of 4.2 % obtained with a close-talking microphone. Moreover, it achieved better recognition performance with a fewer amounts of adaptation data than the conventional MK beamformer.**

I. Introduction

There has been great interest in distant speech recognition (DSR) [5], [6], [7], [8]. In many applications, it is not appropriate to force participants to wear intrusive devices such as close talking microphones. Especially in theme parks, it is perhaps necessary for children to feel free to join attractions without unnatural interface equipment. Accordingly, we address a recognition task of childeren's speech.

In DSR, a target speech signal is corrupted with reverberant and noise effects. Adaptive beamforming is a promising approach to speech enhancement because the distortion of the target signal caused by noise suppression is less than that of single channel processing.

Such adaptive beamformers can be implemented in generalized sidelobe canceler (GSC) configuration [6, §13.37]. The GSC beamformers typically consist of the *quiescent vector*, *blocking matrix* and *active weight vector*. In prior work, we developed GSC beamforming methods which adjust the active weight vector so as to make the beamformer's outputs as super-Gaussian as possible [1], [2], [3], [9]. We demonstrated in [1], [2], [9] that beamforming algorithms based on such criteria can enhance the target speech by using the reflections without *signal cancellation* encountered in the MVDR beamformers. It was also shown that our beamforming methods achieve better recognition performance than a variant of MVDR beamformers. Here, we consider the maximum kurtosis criterion which can be computed in the much simpler way than negentropy. In contrast to BSS techniques, beamforming based on the maximum kurtosis criterion [2], [6] can avoid undesired distortion of the target signal by imposing a distortionless constraint for the look direction. However, in the same as BSS methods, maximum kurtosis beamforming has to resort to a gradient-based numerical optimization algorithm for estimating the active weight vector. Accordingly, it requires a relatively large amount of adaptation data for reliable estimation. Furthermore, the free parameters to be estimated by the gradient method are preferably reduced.

In this work, we take into account the subspace (eigenspace) method [4, §6.8] [10], [11] as a pre-processing step for estimation of the active weight vector. Our motivations behind this idea are to 1) reduce the dimension of the active weight vector and 2) improve speech enhancement performance based on decomposition of the outputs of the blocking matrix into directional and ambient signal components. In order to achieve such decomposition, we perform the eigendecomposition and select the D eigenvectors corresponding to the largest eigenvalues. Such eigenvectors are termed the dominant modes [4, §6.8.3]. The dominant modes are associated with the directional sound sources and the other modes are averaged as a signal model of ambient noise. By doing so, we can readily subtract the averaged ambient noise component from

The authors would like to thank Prof. Jessica Hodgins for giving us the opportunity to study this work. The authors would also like to thank Jill Lehman for conducting the data collection of children speech. Also due thanks are, Wei Chu, Jerry Feng, Ishita Kapur, and Moshe Mahler for their assistance in collecting the speech material used for the experiments described in this work.

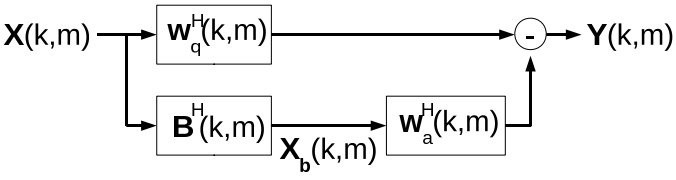

Fig. 1. Schematic of a generalized sidelobe canceling (GSC) beamformer for an active source.

the beamformer's output. Moreover, the reduction of the dimension of the active weight leads to computationally efficient and reliable estimation. Notice that we adjust the active weight vector based on the maximum kurtosis criterion in contrast to the normal dominant-mode rejection (DMR) beamformers [4, §6.8.3] which can be considered as the variant of the MVDR beamformers. Therefore, our technique described here does not suffer from signal cancellation. It is also worth noting that subspace filtering here is analogous to whitening used as a measure of pre-processing in the field of independent component analysis (ICA) [12].

The balance of this paper is organized as follows. Section II describes the conventional MK beamforming algorithm. Section III discusses the subspace method. In the section IV, MK beamforming with the subspace filter is described. Section V shows the speech recognition experiments. We finally describe conclusions and future work in the section VI.

II. Conventional maximum kurtosis beamforming

Consider a subband beamformer in GSC configuration [4, §6.7.3], as shown in Figure 1. The output of a beamformer for a given subband at frame k and frequency bin m can be expressed as

$$Y(k,m) = [\mathbf{w}_{\mathrm{q}}(k,m) - \mathbf{B}(k,m)\mathbf{w}_{\mathrm{a}}(k,m)]^H \mathbf{X}(k,m), \quad (1)$$

where $\mathbf{w}_{\mathrm{q}}(k,m)$ is the *quiescent weight vector* for a source, $\mathbf{B}(k,m)$ is the *blocking matrix*, $\mathbf{w}_{\mathrm{a}}(k,m)$ is the *active weight vector*, and $\mathbf{X}(k,m)$ is the input subband *snapshot vector*. In keeping with the GSC formalism, $\mathbf{w}_{\mathrm{q}}(k,m)$ is chosen to give unity gain in the look direction [6, §13.6]; i.e., to satisfy a distortionless constraint. The blocking matrix $\mathbf{B}(k,m)$ is chosen to be orthogonal to $\mathbf{w}_{\mathrm{q}}(k,m)$, such that $\mathbf{B}^H(k,m)\,\mathbf{w}_{\mathrm{q}}(k,m) = \mathbf{0}$. This orthogonality implies that the distortionless constraint will be satisfied for any choice of $\mathbf{w}_{\mathrm{a}}(k,m)$.

In our study, subband analysis and synthesis are performed with a uniform DFT filter bank based on the modulation of a single prototype impulse response [13][6, §11], which is designed to minimize an indivisual aliasing term.

While the active weight vector is typically chosen to minimize the variance of the beamformer's outputs, we developed an optimization procedure to find that $\mathbf{w}_{\mathrm{a}}(k,m)$ which maximizes kurtosis [2]. The empirical kurtosis of the beamformer's outputs can be expressed as

$$J_m(Y) = E\left[|Y(m)|^4\right] - \beta E\left[|Y(m)|^2\right]^2, \quad (2)$$

where $E[\cdot]$ indicates the expectation operator and β is typically set to $\beta = 3$. The empirical kurtosis (2) measures how *non-Gaussian* Y is [12]. The Gaussian pdf has zero kurtosis; pdfs with positive kurtosis are *super-Gaussian*; those with negative kurtosis are *sub-Gaussian*. Note that the empirical kurtosis measure requires no knowledge of the actual pdf of subband samples of speech, which is its primary advantage over negentropy as a measure of non-Gaussianity. Maximizing the degree of super-Gaussianity yields a weight vector $\mathbf{w}_{\mathrm{a}}$ capable of canceling interference—including incoherent noise that leaks through the sidelobes—without the signal cancellation problems encountered in conventional beamforming.

III. Subspace method

This section describes the subspace method for the output of the blocking matrix in the subband domain. From this section, we omit the frequency index m for the sake of convenience.

In the case that there are neither steering errors nor mismatches between microphones, the blocking matrix's output, $\mathbf{X}_b(k) = \mathbf{B}^H(k)\mathbf{X}(k)$, only contains the directional interference and ambient noise signals. However, in the real environments, it also includes the target signal components due to those errors as well as the reverberant effects.

Let us first denote the D directional signal components contained in the output of the $M_b \times (M_b - 1)$ blocking matrix as

$$\mathbf{V}(k) = [V_1(k), \cdots, V_d(k), \cdots, V_D(k)]^T. \quad (3)$$

Then, the output of the blocking matrix can be expressed as

$$\mathbf{X}_b(k) = \mathbf{A}\mathbf{V}(k) + \mathbf{N}(k) \quad (4)$$

where $\mathbf{A}$ and $\mathbf{N}(k)$ represent the transfer functions and the ambient noise signals, respectively. Notice that the direct path from the target source signal to each microphone is assumed to be excluded from $\mathbf{A}$ because of the distortionless constraint imposed with the blocking matrix.

Assuming that $\mathbf{V}(k)$ and $\mathbf{N}(k)$ are uncorrelated, we can write the covariance matrix of $\mathbf{X}_b$ as

$$\mathbf{\Sigma}_b = E\left[\mathbf{X}_b(k)\mathbf{X}_b^H(k)\right] = \mathbf{A}\mathbf{\Sigma}_v\mathbf{A}^H + \mathbf{\Sigma}_n, \quad (5)$$

where

$$\mathbf{\Sigma}_v = E\left[\mathbf{V}(k)\mathbf{V}^{\mathbf{H}}(k)\right] \text{ and } \mathbf{\Sigma}_n = E\left[\mathbf{N}(k)\mathbf{N}^{\mathbf{H}}(k)\right].$$

The subspace method seeks a set of D linearly independent vectors contained in the subspace, $\Re\{\mathbf{A}\}$, spanned by the column vectors of $\mathbf{A}$. The first step for obtaining such set of the vectors is to solve the generalized eigenvalue (GE) decomposition problem as in [4, §6.8] [10], [11],

$$\mathbf{\Sigma}_v\mathbf{E} = \mathbf{\Sigma}_n\mathbf{E}\mathbf{\Lambda},$$

where $\mathbf{\Lambda}$ is a diagonal matrix of the eigenvalues sorted in the descending order,

$$\mathbf{\Lambda} = \mathrm{diag}\left[\lambda_1, \cdots, \lambda_D, \cdots, \lambda_{M_b-1}\right], \quad (6)$$

and $\mathbf{E}$ is a matrix of the corresponding eigenvectors,

$$\mathbf{E} = \left[\mathbf{e}_1, \cdots, \mathbf{e}_D, \cdots, \mathbf{e}_{M_b-1}\right]. \quad (7)$$

Here, we assume that $\mathbf{\Sigma}_n$ is an identity matrix. Then, we select the eigenvectors with the D largest eigenvalues, $\mathbf{E}_v =$

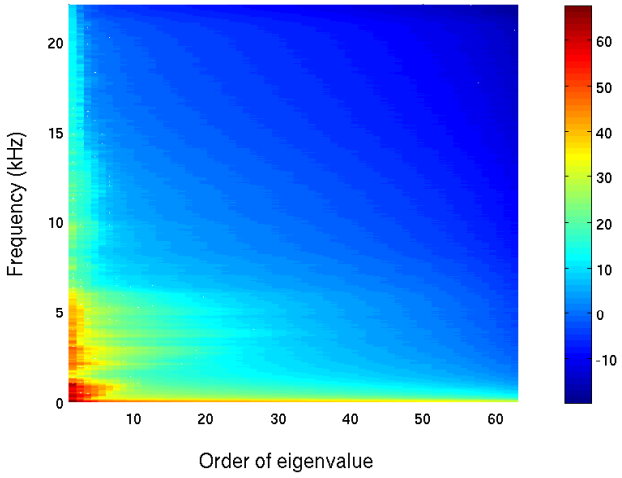

Fig. 2. Three-dimensional representation of the eigenvalue distribution as a function of the order and frequencies.

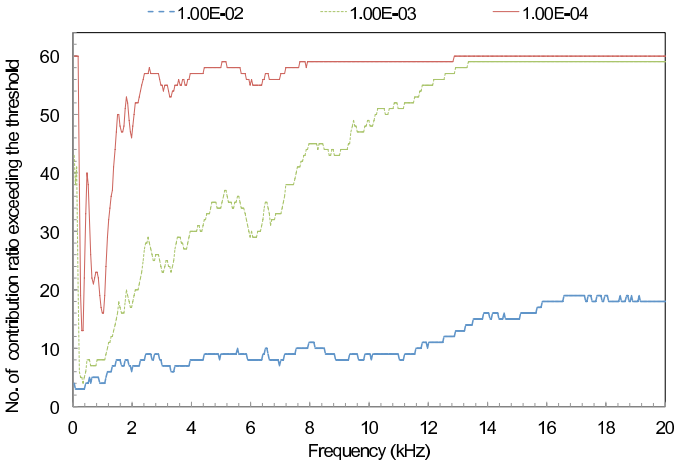

Fig. 3. The average of the numbers of the contribution ratios exceeding the thresholds as a function of frequencies.

$[\mathbf{e}_1, \cdots, \mathbf{e}_D]$. In the similar manner, we define the subspace for the ambient noise as $\mathbf{E}_n = [\mathbf{e}_{D+1}, \cdots, \mathbf{e}_{M_b-1}]$.

The ideal properties of the eigenvectors and eigenvalues can be summarized as follows.

- The subspace spanned by the eigenvectors is equal to that of $\mathbf{A}$, i.e., $\Re\{\mathbf{E}_v\} = \Re\{\mathbf{A}\}$.
- The power of the D directional signals is associated with the D largest eigenvalues.
- The power of $\mathbf{N}(k)$ is equally spread over all the eigenvalues and $M_b - D - 1$ smallest eigenvalues are all equal to σ_N^2, i.e., the noise floor.
- $\Re\{\mathbf{E}_n\}$ is the orthogonal complement of $\Re\{\mathbf{E}_v\}$, i.e., $\Re\{\mathbf{E}_n\} = \Re\{\mathbf{E}_v\} \perp$.

In order to cluster the eigenvectors for the ambient noise, we have to determine the number of the dominant eigenvalues D. Figure 2 illustrates an eigenvalue distribution as a function of the order of the eigenvalues and frequencies. In order to generate the plots of the figures in this section, we computed the eigenvalues from the outputs of the blocking matrix on the real data which will be described in Section V. As shown in Figure 2, it is relatively easy to determine the number of the dominant modes, D, especially in the case that the number of the microphones is much larger than the number of the directional signals. In this work, we determine D based on the threshold of the contribution ratio, $\lambda_i / \sum_{j=1}^{M_b-1} \lambda_j$. Figure 3 shows averages of numbers of the contribution ratios exceeding thresholds, 10^{-2}, 10^{-3} and 10^{-4}, at each frequency. Figure 3 indicates how many dominant modes are used in the lower branch when we ignore the eigenvectors associated with the lower contribution ratio than the threshold. It is clear from Figure 3 that the lower threshold for the contribution ratio is set, the more eigenvectors are used.

Ideally, the number of the dominant modes should be equal to the number of the directional sound sources over all the frequencies. However, by closely looking at Figure 3, we can observe that the number of the dominant modes depends on frequencies. In particular, more eigenvectors tends to be chosen in high frequencies due to absence of directional signals with high frequency components. In the case that the strong coherent signals do not exist, the distribution of the eigenvalues become homogeneous, which results in the uniform contribution ratios.

In optimization of the active weight vector, we estimate each component corresponding to the ambient noise signal separately. Accordingly, we use the sum of the eigenvectors for the ambient noise space as $\tilde{\mathbf{e}}_n = \sum_{d=D+1}^{M_b-1} \mathbf{e}_d$. Our subspace filter can be now written as

$$\mathbf{U} = [\mathbf{e}_1, \cdots, \mathbf{e}_D, \tilde{\mathbf{e}}_n,]\,. \tag{8}$$

Note that we assume the covariance matrix in (5) can be approximated as

$$\mathbf{\Sigma}_b \approx \sum_{d=1}^{D} \lambda_d \mathbf{e}_d \mathbf{e}_d^H + \bar{\sigma}_N^2 \tilde{\mathbf{e}}_n \tilde{\mathbf{e}}_n^H\,, \tag{9}$$

where

$$\bar{\sigma}_N^2 = \frac{1}{M_b - D - 1} \sum_{d=D+1}^{M_b-1} \lambda_d$$

With the outputs of the subspace filter, we estimate the active weight vector providing the maximum kurtosis value. If the output of the subspace filter is a noise signal, the corresponding component of the active weight vector should be adjusted so as to subtract the noise component from the output of the quiescent vector. If it is an echo of the target signal, the active weight vector should shift the phase and add the component to the target signal in order to strengthen it. These operations would be easier by separating the echo from the ambient noise component with the subspace filter.

IV. Maximum Kurtosis Beamforming with Subspace Filtering

Figure 4 shows configuration of our new MK beamformer. Our beamformer's output can be expressed as

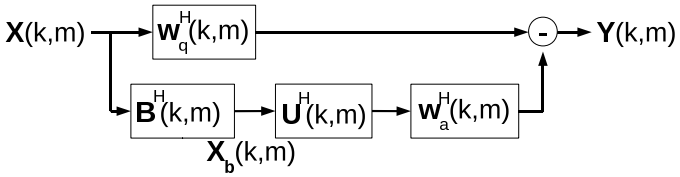

Fig. 4. Maximum kurtosis beamformer with the subspace filter.

$$Y(k) = [\mathbf{w}_{\mathrm{q}}(k) - \mathbf{B}(k)\mathbf{U}(k)\mathbf{w}_{\mathrm{a}}(k)]^H \ \mathbf{X}(k). \tag{10}$$

The active weight vector is adjusted so as to achieve the maximum kurtosis of beamformer's output. Zelinski post-filtering can then be performed on the output of the beamformer [14].

The difference between (1) and (10) is the subspace filter between the blocking matrix and active weight vector. Our subspace filter can decompose the output vector into the directional signal and ambient noise components. Therefore, we only need to estimate the phase shifts of the active weight vector on the constrained subspace fundamentally [12, §7.4]. Moreover, the solution of the general eigenvector decomposition is less dependent of the initial values than that of the gradient algorithm for multi-dimensional maximization.

A. Block-Wise Adaptation of the Active Weight Vector

Based on equation (10), the kurtosis of the outputs is computed from a block of input subband samples at each block instead of using the entire utterance data. We incrementally update the dominant modes and active weight vector at each block b consisting of L_b samples here. Accordingly, the beamformer's output of (10) should be precisely re-rewritten as

$$Y(k) = [\mathbf{w}_{\mathrm{q}}(k) - \mathbf{B}(k)\mathbf{U}(\lfloor k/L_b \rfloor)\mathbf{w}_{\mathrm{a}}(\lfloor k/L_b \rfloor)]^H \ \mathbf{X}(k). \tag{11}$$

where $\lfloor . \rfloor$ is the floor function and $\lfloor k/L_b \rfloor$ indicates the block index b. The kurtosis for a block of L_b samples starting from frame b_s can be expressed as

$$J_b(Y) = \left(\frac{1}{L_b} \sum_{k=b_s}^{b_s+L_b-1} |Y(k)|^4 \right) - \beta \left(\frac{1}{L_b} \sum_{k=b_s}^{b_s+L_b-1} |Y(k)|^2 \right)^2 . \tag{12}$$

where b_s is zero at the first input sample and shifted with L_b after one block is processed.

In order to improve robustness by inhibiting the formation of excessively large sidelobes, we apply a regularization term [6] to the cost function (12) and have the modified optimization criterion

$$\mathcal{J}_b(Y;\alpha) = J_b(Y) - \alpha \|\mathbf{w}_{\mathrm{a}}(b)\|^2 \tag{13}$$

where we set α = 0.01 based on the results of the speech recognition experiments in prior work [1], [2]. In addition to the regularization term, we also impose a unity constraint on a norm of the active weight vector so as to prevent it from exceeding that of the quiescent vector.

We estimate the active weight vector which maximizes the sum of the kurtosis and regularization term (13) under the norm constraint at each block . In the absence of a closed-form solution, we resorted to the *gradient descent algorithm* [15, §1.6]. Upon substituting (11) into (13) and taking the partial derivative with respect to the active weight vector, we obtain

$$\begin{aligned} \frac{\partial \mathcal{J}_b(Y;\alpha)}{\partial \mathbf{w}_{\mathrm{a}}(b)^*} &= -\frac{2}{L_b} \left(\sum_{k=b_s}^{b_s+L_b-1} |Y(k)|^2 \mathbf{U}^H(b)\mathbf{B}^H(k)\mathbf{X}(k)Y^*(k) \right) \\ &+ \frac{2\beta}{L_b^2} \left(\sum_{k=b_s}^{b_s+L_b-1} |Y(k)|^2 \right) \left(\sum_{k=b_s}^{b_s+L_b-1} \mathbf{U}^H(b)\mathbf{B}^H(k)\mathbf{X}(k)Y^*(k) \right) \\ &- \alpha \mathbf{w}_{\mathrm{a}}(b). \end{aligned} \tag{14}$$

The gradient (14) is iteratively calculated with a block of subband samples until the kurtosis value of the beamformer's outputs converges. For the gradient algorithm, the active weight vectors are initialized with the estimates at the previous block. The active weight vector of the first block is initialized with $\mathbf{w}_{\mathrm{a}} = [0, \cdots, 1]^T$ because the last component corresponds to the ambient noise which should be subtracted from the output of the quiescent vector.

The beamforming algorithm can be summarized as follows:

1) Initialize the active weight with $\mathbf{w}_{\mathrm{a}}(0) = [0, \cdots, 1]^T$.
2) Given estimates of time delays, calculate the quiescent vector and blocking matrix.
3) For each block of input subband samples, recursively update the covariance matrix as $\mathbf{\Sigma}_b(b) = \mu\mathbf{\Sigma}_b(b-1) + (1-\mu)\mathbf{\Sigma}_b(b)$ where μ is the forgetting factor, calculate the dominant modes $\mathbf{U}^H(b)$ and estimate the active weight vector $\mathbf{w}_{\mathrm{a}}(b)$ based on the gradient information computed with (14) subject to the norm constraint until the kurtosis value of the beamformer's outputs converges.
4) Initialize the active weight vector obtained in step 3 for the next block and go to the step 2.

Our preliminary experiments revealed that this block-wise method is able to track a non-stationary sound source, and provides a more accurate gradient estimate than *sample-by-sample* gradient estimation algorithms.

V. Speech recognition experiment

We ran speech recognition experiments on children's speech data captured with a microphone array. This section describes our experimental conditions and the results.

Figure 5 shows a flow chart of the distant speech recognition (DSR) system used in our experiments. Our DSR system first estimates the time delays based on the phase transform (PHAT) [6, §10.1]. Then, reliable channels are selected based on the maximum multi-channel cross coefficient criterion(MCCC) [16]. Following channel selection, beamforming and post-filtering are performed. The enhanced speech are input to our automatic speech recognition (ASR) system.

Our basic automatic speech recognition (ASR) system was trained on two publicly available corpora of children's speech:

1) the Carnegie Mellon University (CMU) Kids' Corpus, which contains 9.1 hours of speech from 76 speakers;

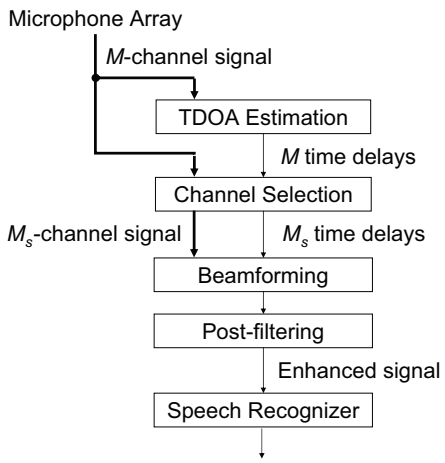

Fig. 5. A flow chart of our distant speech recognition system.

2) the Center for Speech and Language Understanding (CSLU) Kids' Corpus, which contains 4.9 hours of speech from 174 speakers.

The feature extraction used for the experiments was based on cepstral features estimated with a warped *minimum variance distortionless response* (MVDR) spectral envelope of model order 30 [6, §5.3]. Front-end analysis involved extracting 20 cepstral coefficients per frame of speech, and then performing *cepstral mean normalization* (CMN). The final features were obtained by concatenating 15 consecutive frames of cepstral coefficients together, then performing *linear discriminant analysis* (LDA), to obtain a feature of length 42. The LDA transformation was followed by a second CMN step, then a global semi-tied covariance transform estimated with a maximum likelihood criterion [17].

The acoustic HMM was initialized from a context independent model with three states per phone with the global mean and variance of the training data. Thereafter, five iterations of Viterbi training [6, §8.1.5] were conducted. This was followed by an additional five iterations whereby optional silences and optional breath phones were allowed between words. The next step was to treat all triphones as distinct and train three-state single-Gaussian models for each in order to cluster the states [18]. In the final stage of conventional training, the context-dependent state-clustered model was initialized with a single Gaussian per codebook from the context-independent model; three iterations of Viterbi training followed by splitting the Gaussian with the model training steps. These steps were repeated until no more Gaussians had sufficient training counts to allow for splitting. The conventional model had 1,200 states and a total of 25,702 Gaussian components. Conventional training was followed by *speaker-adapted training* (SAT) as described in [6, §8.1.3].

In our experiments, the ASR system consisted of three passes:

1) Recognize with the unadapted acoustic model;
2) Estimate *vocal tract length normalization* (VTLN) [6, §9.1], *maximum likelihood linear regression* (MLLR) [6, §9] and *constrained maximum likelihood linear regression* (CMLLR) [6, §9] parameters, then recognize once more with the adapted conventially trained model;
3) Estimate VTLN, MLLR and CMLLR parameters for the SAT model, then recognize with same.

For all but the first unadapted pass, unsupervised speaker adaptation was performed based on word lattices from the previous pass.

Test data for our experiments were collected at the Carnegie Mellon University Children's School. The speech material in this corpus was captured with a 64-channel Mark IV microphone array; the elements of the Mark IV were arranged linearly with a 2 cm intersensor spacing. In order to provide a reference for the DSR experiments, the subjects of the study were also equipped with Shure lavelier microphones with a wireless connection to a preamp input. All the audio data were captured at 44.1 kHz. The test set consists of 356 (1305 words) utterances spoken by an adult and 354 phrases (1,297 words) uttered by nine children. The children were native-English speakers (aged four to six). They were asked to play *Copycat*, a listen-and-repeat paradigm in which the adult experimenter speaks a phrase and the child tries to copy both pronunciation and intonation. As is typical for children in this age group, pronunciation was quite variable and the words themselves sometimes indistinct.

The search graph for ASR was created by constructing a finite-state automaton by stringing *Copycat* utterances in parallel between a start and end state. This acceptor was convolved together with a finite-state transducer representing the phonetic transcriptions of the 147 words in the *Copycat* vocabulary. Thereafter this transducer was convolved with the HC transducer representing the context-dependency decision tree estimated during state-clustering [6, §7.3.4].

Table I shows word error rates (WERs) of every decoding pass obtained with the single distant microphone (SDM), super-directive beamforming (SD BF), conventional maximum kurtosis beamforming (MK BF) and maximum kurtosis beamforming with the subspace filter (MK BF w SF). The WERs obtained with the lapel microphone are also described as a reference . It is clear from the table I that the speaker adaptation techniques can significantly reduce the WERs. It is also clear from table I that the new maximum kurtosis beamformer achieved the best recognition performance after the third pass.

Table II shows the WERs for the threshold of the contribution ratio. The difference of the threshold does not give a big impact on recognition performance. However, the computational reduction can be significant especially in the case that the number of the channels are large because we can reduce the dimension of the active weight vector without sacrificing recognition performance by setting a low threshold value.

Table III shows the WERs of the conventional and new MK beamforming algorithms as a function of amounts of adaptation data at each block. We can see from table III that MK beamforming with subspace filtering (MK BF w SF)

Algorithm	Pass 1		Pass 2		Pass 3	
	Exp.	Child	Exp.	Child	Exp.	Child
SDM	9.2%	31.0%	3.8%	17.8%	3.4%	14.2%
SD BF	5.4%	24.4%	2.5%	9.6%	2.2%	7.6%
MK BF	5.4%	25.1%	2.5%	9.0%	2.1%	6.5%
MK BF w SF	6.3%	25.4%	1.2%	7.4%	0.6%	5.3%
CTM	3.0%	12.5%	2.0%	5.7%	1.9%	4.2%

TABLE I
WORD ERROR RATES (WERS) FOR EACH DECODING PASS.

Threshold for the contribution ratio	Pass 2		Pass 3	
	Exp.	Child	Exp.	Child
10^{-1}	0.8%	8.3%	0.5%	6.3%
10^{-2}	1.2%	7.4%	0.6%	5.3%
10^{-3}	1.3%	8.7 %	1.2%	5.9%

TABLE II
WERS FOR THE THRESHOLDS OF THE CONTRIBUTION RATIO.

Algorithm	Block size (second)	Pass 2		Pass 3	
		Exp.	Child	Exp.	Child
Conventional MK BF	0.25	4.4%	15.8%	3.5%	12.0%
	0.5	3.4%	9.2%	3.1%	7.3%
	1.0	2.4%	10.3%	2.2%	6.9%
	2.5	2.5%	9.0%	2.1%	6.5%
MK BF w SF	0.25	2.5%	14.1%	1.5%	9.7%
	0.5	1.3%	8.7%	1.0%	7.0%
	1.0	1.2%	7.4%	0.6%	5.3%

TABLE III
WERS AS A FUNCTION OF AMOUNTS OF ADAPTATION DATA.

provides better recognition performance with the same amount of the data than conventional MK beamforming. In the case that a few amount of data are available, the solution of the active weight vector obtained by normal MK beamforming does not always improve recognition performance due to the dependency of the initial value and noisy gradient information which can significantly change over the blocks. The results in Table III suggest that unreliable estimation of the active weight vector can be avoided by constraining the search space on the subspace spanned by the dominant modes and basis vector representing the ambient noise component. Note that the solution of the eigendecomposition does not depend on the initial value in contrast to the gradient-based numerical optimization algorithm.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed the cascade of the subspace filter and blocking matrix for the maximum kurtosis beamformer. We also demonstrated through a set of the DSR recognition results that the beamforming algorithm proposed here effectively suppresses interference and ambient noise signals.

We plan to use different criteria for the detection of the subspace dimension such as AIC and MDL measures [4, §7.8]. We also plan to apply the online subspace learning algorithms [4, §7.9][19] in order to further reduce computation.

REFERENCES

[1] K. Kumatani, J. McDonough, D. Klakow, P. N. Garner, and W. Li, "Adaptive beamforming with a maximum negentropy criterion," *IEEE Trans. Audio, Speech and Language Processing*, August 2008.

[2] 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, September 2008.

[3] K. Kumatani, J. McDonough, B. Rauch, and D. Klakow, "Maximum negentropy beamforming using complex generalized gaussian distribution model," in *Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, Pacific Grove, CA, USA, 2010.

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

[5] H. K. Maganti, D. Gatica-Perez, and I. McCowan, "Speech enhancement and recognition in meetings with an audio-visual sensor array," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 2257–2269, 2007.

[6] M. Wölfel and J. McDonough, *Distant Speech Recognition.* New York: Wiley, 2009.

[7] E. Zwyssig, M. Lincoln, and S. Renals, "A digital microphone array for distant speech recognition," in *Proc. of ICASSP*, Dallas, Texas, USA, 2010.

[8] S. Araki, T. Hori, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, M. Delcroix, and K. Kinoshita, "Low-latency meeing recognition and understanding using distant microphones," in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Edinburgh, UK, 2011.

[9] K. Kumatani, T. Gehrig, U. Mayer, E. Stoimenov, J. McDonough, and M. Wölfel, "Adaptive beamforming with a minimum mutual information criterion," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, pp. 2527–2541, 2007.

[10] R. Roy and T. Kailath, "Esprit–estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, pp. 984–995, 1989.

[11] F. Asano, S. Ikeda, M. Ogawa, H. Aso, and N. Kitawaki, "Combined approach of array processing and independent component analysis for blind separation of acoustic signals," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 3, pp. 204–215, May 2003.

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

[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] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with post-filtering," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 240–259, 1998.

[15] D. P. Bertsekas, *Nonlinear Programming.* Belmont, MA, USA: Athena Scientific, 1995.

[16] K. Kumatani, J. McDonough, J. Lehman, and B. Raj, "Channel selection based on multichannel cross-correlation coefficients for distant speech recognition," in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Edinburgh, UK, 2011.

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

[18] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree–based state tying for high accuracy acoustic modelling," in *Proc. of HLT*, Plainsboro, NJ, USA, 1994, pp. 307–312.

[19] R. Badeau, B. David, and G. Richard, "Fast approximated power iteration subspace tracking," *IEEE Trans. on Signal Processing*, vol. 53, pp. 2931–2941, 2005.