BLOCK-WISE INCREMENTAL ADAPTATION ALGORITHM FOR MAXIMUM KURTOSIS BEAMFORMING

Kenichi Kumatani[*]

Disney Research, Pittsburgh
4720 Forbes Avenue, Lower Level, Suite 110
Pittsburgh, PA 15213, USA
k_kumatani@ieee.org

John McDonough, Bhiksha Raj

Carnegie Mellon University
Gates-Hillman Complex, 5000 Forbes Avenue
Pittsburgh, PA 15213, USA
{johnmcd,bhiksha}@cs.cmu.edu

ABSTRACT

In prior work, the current authors investigated beamforming algorithms that exploit the non-Gaussianity of human speech. The beamformers proposed in [1, 2, 3] are designed to maximize the kurtosis or negentropy of the subband output subject to the distortionless constraint for the direction of interest. Such techniques are able to suppress interference signals as well as reverberation effects without *signal cancellation*. They require, however, multiple passes of processing for each utterance in order to estimate the *active weight vector*. Hence, they are unsuitable for online implementation. In this work, we propose an online implementation of the maximum kurtosis beamformer. In a set of distant speech recognition experiments on far-field data, we demonstrate the effectiveness of the proposed technique. Compared to a single channel of the array, the proposed algorithm reduced word error rate from 15.4% to 6.5%.

Index Terms— Kurtosis, Beamforming, Microphone array, Distant speech recognition

1. INTRODUCTION

In prior work, the current authors investigated the use of optimization criteria for beamforming that exploit the non-Gaussianity of human speech. This non-Gaussianity is a characteristic that can be easily exploited in beamformer design, in that clean speech is highly super-Gaussian, but becomes more nearly Gaussian when corrupted by noise or reverberation [4, §13.5.2]. The algorithms examined in our prior work were designed to either maximize kurtosis [1] or negentropy [2, 3] of the subband output of a generalized sidelobe canceller (GSC) [4, §13]. It was also shown in [1, 2] that those beamforming techniques can effectively remove noise and reverberation effects without the signal cancellation problems encountered in the conventional algorithms based on second-order statistics such as minimum variance distortionless response (MVDR) beamformers [4, §13.3]. However, those methods required making multiple passes through the data and hence are unsuitable for online implementation. Thus, in this work, we propose an online implementation of the maximum kurtosis beamformer, wherein the active weights of the GSC are updated for each block of samples after a single pass through each utterance.

Distant speech recognition (DSR) has been of interest for interactive speech applications. DSR can be especially useful for young children who may find CTMs too cumbersome and intrusive to use in interactive attractions. Moreover, the accuracy of the state of the art automatic speech recognition (ASR) systems is perhaps a good objective measure for the speech intelligibility because it takes into account the stochastic distance between correct and incorrect phoneme sequences with acoustic models trained with many hours of clean speech data and information for discriminating phonemes is only used in ASR. In this work, we combine these heretofore separate research tracks, by testing the algorithms mentioned above on speech data collected with lapel microphones and a linear microphone array with 64 sensors. The subjects of the data collection were children aged 4 to 6. In a set of DSR experiments on the speech material, we show the effectiveness of the proposed technique.

The balance of this work is organized as follows. In Section 2, we briefly review the non-Gaussian characteristics of speech, along with how these basic characteristics can be successfully exploited in developing effective beamforming algorithms. Our experimental results are presented in Section 3, as well as a discussion thereof. Finally, in Section 4 we draw our conclusions about this work and outline our plans for the future.

2. MAXIMUM KURTOSIS BEAMFORMING

2.1. Super-Gaussianity and Kurtosis

The central limit theorem states that the sum of independent random variables (r.v.s) is approximately Gaussian-distributed in the limit as more components are added regardless of the probability density functions (pdfs) of the individual components. It is also known that the distribution of information-bearing signals such as clean speech is not Gaussian. In fact, the actual distribution of the clean speech signals fits in a super-Gaussian pdf which is characterized by peaky and heavy-tailed probability mass distribution [4, §13.5.2]. Therefore, speech can be enhanced by adjusting a beamformer's weights so as to make the outputs as super-Gaussian as possible. The kurtosis is one of the popular criteria to measure the degree of non-Gaussianty, that is, how far the distribution of r.v.s is from Gaussian. The *excess kurtosis* or simply *kurtosis* of a r.v. Y with zero mean, can be expressed as

$$\text{kurt}(Y) \triangleq \mathcal{E}\{Y^4\} - \beta(\mathcal{E}\{Y^2\})^2, \tag{1}$$

[*]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.

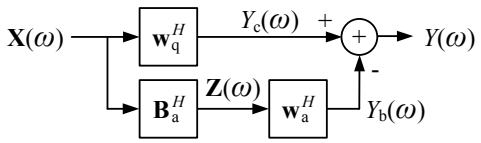

Figure 1: Configuration of the generalized sidelobe canceller (GSC).

where β is a positive constant, which is typically set to three in order to ensure that 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 can be computed without knowledge of the actual pdf of subband samples of speech, which is its primary advantage over other measures of non-Gaussianity. However, the empirical kurtosis can be greatly influenced by a few samples with a low observation probability; Hyvärinen and Oja [5] note that negentropy is generally more robust in the presence of outliers than kurtosis.

2.2. Generalized Sidelobe Canceller Beamforming

Consider a subband beamformer in the generalized sidelobe canceller (GSC) configuration [4, §13.6] shown in Figure 1. Let us first denote the input subband *snapshot vector* at frame k as $\mathbf{X}(k)$ where the frequency index is omitted. The output of a beamformer at frame k can be expressed as

$$Y(k) = [\mathbf{w}_{\rm q}(k) - \mathbf{B}(k)\mathbf{w}_{\rm a}(k)]^H \ \mathbf{X}(k), \tag{2}$$

where $\mathbf{w}_{\rm q}(k)$ is the *quiescent weight vector* for a source, $\mathbf{B}(k)$ is the *blocking matrix*, $\mathbf{w}_{\rm a}(k)$ is the *active weight vector*. In this work, we suppress indices for frquency bins. In keeping with the GSC formalism, $\mathbf{w}_{\rm a}(k)$ is chosen to give unity gain in the desired or *look direction* [4, §13.6]; i.e., to satisfy a *distortionless constraint*. The blocking matrix $\mathbf{B}(k)$ is chosen to be orthogonal to $\mathbf{w}_{\rm q}(k)$, such that $\mathbf{B}^H(k)\,\mathbf{w}_{\rm q}(k) = \mathbf{0}$. This orthogonality implies that the distortionless constraint will be satisfied for any choice of $\mathbf{w}_{\rm a}(k)$.

While the active weight vector is typically chosen to minimize the variance of the beamformer's outputs, here we develop on online optimization procedure to find that $\mathbf{w}_{\rm a}(k)$ which maximizes kurtosis. Maximizing the degree of super-Gaussianity yields the active weight weight vector capable of canceling interference—including incoherent noise that leaks through the sidelobes—without the signal cancellation problems encountered in the MVDR beamformers.

We perform subband analysis and synthesis processing with a uniform DFT filter bank based on the modulation of a single prototype impulse response [4, §11], which was designed to minimize each aliasing term individually [6]. For experiments described in section 3, we used the filter prototype with 1024 subbands, the length of the prototype was 2048 and the decimation factor which corresponds to the frame shift was 128. Those values are chosen based on our prior work [1, 2, 3].

2.3. Estimation of the Active Weight Vector

In [1], the kurtosis of the beamformer's output was computed over an entire utterance. While such an algorithm is feasible for relatively small arrays of eight elements, it becomes computationally intractable for the large array considered in this work. It also has an unacceptably slow response when a long utterance must be processed.

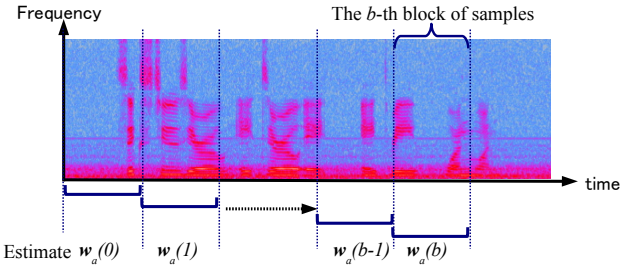

Figure 2: Visualization of the block-wise estimation of the active weight vector.

In our prior work, we minimized kurtosis of the beamformer's output computed with all the incoming snapshots consisting of K frames. Such a cost function can be written as

$$J(Y) = \left(\frac{1}{K}\sum_{k=0}^{K-1}|Y(k)|^4\right) - \beta\left(\frac{1}{K}\sum_{k=0}^{K-1}|Y(k)|^2\right)^2. \tag{3}$$

Clearly, the computational amount becomes significant as the number of the input snapshots K is large. Hence, we incrementally update the active weight vector at each block b consisting of L_b samples here. Accordingly, the beamformer's output of (2) should be precisely re-rewritten as

$$Y(k) = [\mathbf{w}_{\rm q}(k) - \mathbf{B}(k)\mathbf{w}_{\rm a}(\lfloor k/L_b \rfloor)]^H \ \mathbf{X}(k). \tag{4}$$

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

$$J_b(Y) = \left(\frac{1}{L_b}\sum_{k=b_s}^{b_s+L_b-1}|Y(k)|^4\right) - \beta\left(\frac{1}{L_b}\sum_{k=b_s}^{b_s+L_b-1}|Y(k)|^2\right)^2. \tag{5}$$

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 [4] to the cost function (5) and have the modified optimization criterion

$$\mathcal{J}_b(Y;\alpha) = J_b(Y) - \alpha\mathcal{E}\{\|\mathbf{w}_{\rm a}(b)\|^2\} \tag{6}$$

where we set α = 0.1 based on the results of the speech recognition experiments in prior work [1, 2, 3]. In addition to the regularization term, we also impose a 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 (6) under the norm constraint at each block . Figure 2 visualizes with a spectrogram that the active weight vector is estimated with each block of the subband samples.

In the absence of a closed-form solution, we resorted to the *gradient descent algorithm* [7, §1.6]. Upon substituting (5) into (6) and taking the partial derivative with respect to the active weight vector, we obtain

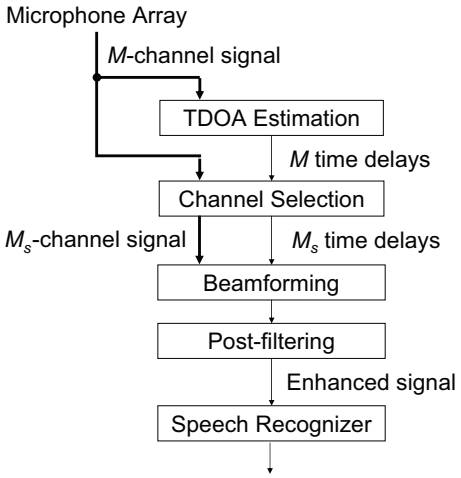

Figure 3: A flow chart of our distant speech recognition system.

$$\frac{\partial \mathcal{J}_b(Y;\alpha)}{\partial \mathbf{w}_{\mathrm{a}}{}^*(b)} = -2\left(\frac{1}{L_b}\sum_{k=b_s}^{b_s+L_b-1}|Y(k)|^2\mathbf{B}^H(k)\mathbf{X}(k)Y^*(k)\right) + 2\beta\left(\frac{1}{L_b}\sum_{k=b_s}^{b_s+L_b-1}|Y(k)|^2\right)\left(\frac{1}{L_b}\sum_{k=b_s}^{b_s+L_b-1}\mathbf{B}^H(k)\mathbf{X}(k)Y^*(k)\right) - \alpha\mathbf{w}_{\mathrm{a}}(b). \tag{7}$$

The gradient (7) is repeatedly calculated with a block of subband samples until it converges. For the gradient algorithm, the active weight vectors are initialized with the estimates at the previous block; the first block is initialized with active weights of zero. Our preliminary experiments revealed that this block-wise batch method is able to track a non-stationary sound source, and provides a more accurate gradient estimate than sample-by-sample update algorithms.

The beamforming algorithm can be summarized as follows:

1. Initialize the active weight with $\mathbf{w}_{\mathrm{a}}(0) = \mathbf{0}$.
2. Given estimates of time delays, calculate the quiescent vector and blocking matrix.
3. For each block of input subband samples, $b = 1, 2, \cdots$, repeat estimation of the active weight vector $\mathbf{w}_{\mathrm{a}}(b)$ based on the gradient information computed with (7) subject to the norm constraint until it converges.
4. set the initial active weight vector for the next block with the current estimate as $\mathbf{w}_{\mathrm{a}}(b+1) \Leftarrow \mathbf{w}_{\mathrm{a}}(b)$ and go to the step 2.

3. EXPERIMENTS

Figure 3 shows a flow chart of the distant speech recognition (DSR) system used in our experiments. Our DSR system involves the time delay estimation based on the phase transformation [4, §10.1]. Then, the channels for beamforming are selected based on the method described in [8]. Following beamforming, Zelinski post-filtering [9], a variant of Wiener filtering, is carried out in order to remove the noise uncorrelated among the sensors.

We use a weighted finite-state transducer (WFST) decoder [4, §7.2] for speech recognition experiments. The feature extraction used for the ASR experiments reported here was based on cepstral features estimated with a warped *minimum variance distortionless response* (MVDR) spectral envelope of model order 30 [4, §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 [10]. The details of training procedures for acoustic models are described in [8].

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

1. Recognize with the unadapted conventionally trained model;
2. Estimate *vocal tract length normalization* (VTLN) [11], *maximum likelihood linear regression* (MLLR) [12] and *constrained maximum likelihood linear regression* (CMLLR) [13] parameters, then recognize once more with the adapted conventionally trained model;
3. Estimate VTLN, MLLR and CMLLR parameters for the SAT model, then recognize with the adapted model.

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

3.1. Recognition results

Test data for 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. This was required to enable voice prompt suppression experiments. All the audio data were captured at 44.1 kHz with a 24-bit per sample resolution.

The test set consists of 354 utterances (1,297 words) spoken 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 an 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 the recognition experiments was created by initially 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 [4, §7.3.4].

Table 1 shows word error rates (WERs) of every decoding pass obtained with one of 64 microphones, super-directive (SD) beamforming, maximum kurtosis (MK) beamforming and lapel microphone. Here, the active weight vectors of the MK beamformer were incrementally adapted with 2.5 seconds of block data. In Table 1, the numbers of channels is automatically determined by the discriminant method [14]. The number of channels used for beamforming ranged from 33 to 62 and the average was 45.

Table 1 demonstrates that the improvement from the adaptation techniques is dramatic. The reduction in the WER from the first

Algorithm	Pass (%WER)		
	1	2	3
Single distant microphone	38.1	19.8	15.4
SD beamforming	24.4	9.6	7.6
MK beamforming	25.1	9.0	6.5
Lapel microphone	19.3	5.9	5.2

Table 1: Word error rates (WERs) for children's speech at each decoding pass.

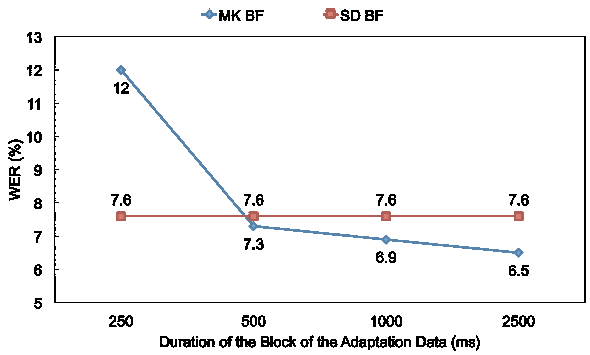

Figure 4: WER after the third recognition pass as a function of the duration of the adaptation data.

pass to the third is approximately four-fold in the case of MK beamforming. It is also clear that the performance of far-field speech recognition can be improved by beamforming techniques, and the MK beamforming algorithm achieves the best performance in the experiments. The MK beamforming technique provides almost the same recognition performance as the lapel microphone. In contrast to the MVDR beamformers, the MK beamformer can be adapted when the target signal is present.

Figure 4 shows WERs of the MK beamformer as a function of the duration of the block of data used for adaptation. In Figure 4, the WER of the SD beamformer is illustrated as a reference. It is clear from Figure 4 that the larger the amount of the block data is, the better recognition performance is. It is also clear from Figure 4 that the WERs of the MK beamformer are higher than those obtained with SD beamforming when the block size is smaller than 500 milliseconds. These results suggest that stable estimation of the active weight vectors requires a certain amount of the block data.

4. CONCLUSIONS

In this work we have proposed the online maximum kurtosis beamformer. We have demonstrated that this algorithm, while remaining computationally tractable, provides superior performance to the super-directive design. We also analyzed how much data is required for stable estimation of the active weight vector.

We plan to develop the estimation method of the active weight vector with a variable block length in order to find short, stationary segments. We also plan to combine conventional beamforming algorithms to enhance the convergence speed. Finally, we plan to integrate beamforming more tightly with the voice prompt suppression algorithms discussed in [15].

5. REFERENCES

[1] 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.

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

[3] K. Kumatani, B. Rauch, J. McDonough, and D. Klakow, "Maximum negentropy beamforming using complex generalized gaussian distribution model," in *Proc. Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, 2010, pp. 1420 –1424.

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

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

[6] 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. ICASSP*, 2008.

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

[8] K. Kumatani, J. McDonough, J. Lehman, and B. Raj, "Channel selection based on multichannel cross-correlation coefficients for distant speech recognition," *Proc. Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, May 2011.

[9] 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.

[10] M. J. F. Gales, "Semi–tied covariance matrices for hidden Markov models," *IEEE Trans. on SAP*, vol. 7, pp. 272–281, 1999.

[11] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. of ICASSP*, vol. I, 1996, pp. 346–8.

[12] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Jour. on CSL*, vol. 9, no. 2, pp. 171–185, 1995.

[13] M. J. F. Gales, "Maximum likelihood linear transformations for HMM–based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[14] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. on Systems, Man, and Cybernetic*, vol. SMC-9, pp. 62–66, 1979.

[15] J. McDonough, K. Kumatani, and B. Raj, "On the combination of voice prompt suppression with maximum kurtosis beamforming," in *Proc. WASPAA*, submitted for publication, 2011.