Maximum Kurtosis Beamforming with the Generalized Sidelobe Canceller

Kenichi Kumatani[1], *John McDonough*[2], *Barbara Rauch*[2], *Philip N. Garner*[1], *Weifeng Li*[1], *John Dines*[1]

[1]IDIAP Research Institute, Martigny, Switzerland
[2]Spoken Language Systems at Saarland University, Saarbrücken, Germany
k_kumatani@ieee.org john.mcdonough@lsv.uni-saarland.de

Abstract

This paper presents an adaptive beamforming application based on the capture of far-field speech data from a real single speaker in a real meeting room. After the position of a speaker is estimated by a speaker tracking system, we construct a subband-domain beamformer in generalized sidelobe canceller (GSC) configuration. In contrast to conventional practice, we then optimize the active weight vectors of the GSC so that the distribution of an output signal is as non-Gaussian as possible. We consider kurtosis in order to measure the degree of non-Gaussianity. Our beamforming algorithms can suppress noise and reverberation without the signal cancellation problems encountered in conventional beamforming algorithms. We demonstrate the effectiveness of our proposed techniques through a series of far-field automatic speech recognition experiments on the Multi-Channel Wall Street Journal Audio Visual Corpus (MC-WSJ-AV). The beamforming algorithm proposed here achieved a 13.6% WER, whereas the simple delay-and-sum beamformer provided a WER of 17.8%.

Index Terms: far-field speech recognition, microphone array, beamforming

1. Introduction

There has been great and growing interest in microphone array processing for hands-free speech recognition [1, 2]. Such techniques 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. 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 [3, §13.6]. Subject to this *distortionless 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* [4]. To avoid the latter, the adaptation of the active weight vector is typically halted whenever the desired source is active.

In [5], we proposed a new beamforming algorithm which adjusted the active weight vectors to maximize the *negentropy* of the beamformer's outputs. Negentropy indicates how far a probability density function (pdf) of a particular signal is from Gaussian. In other words, it represents the degree of super-Gaussianity of a distribution [6]. The pdf of speech is in fact super-Gaussian [2, 7], but it becomes closer to Gaussian when the speech is corrupted by noise or reverberation. Hence, noise and reverberation can be suppressed by adjusting the active weight vector of the GSC to provide a signal with the highest possible negentropy. We also demonstrated in [5] that maximum negentropy (MN) beamforming is free from the signal cancellation problem and provides the better recognition performance than conventional methods.

In this work, we consider *kurtosis* as a criterion for estimating the active weight vectors in a GSC. The kurtosis also measures the degree of super-Gaussianity of a pdf [6]. We optimize the active weight vectors of a GSC so as to achieve the output with the *maximum kurtosis* (MK). After beamforming, *Zelinski* post-filtering is performed to further enhance the speech by removing residual noise [8]. Much like the MN beamformer, the MK beamformer can suppress noise and reverberation without the signal cancellation problem encountered in conventional adaptive beamforming algorithms. In contrast to negentropy, kurtosis does not require knowledge of the actual pdf of subband samples of speech. Rather, kurtosis can be simply calculated in a non-parametric manner. However, the kurtosis measure is influenced by samples with a low observation probability [6]. It is worth mentioning that Gillespie et al. [9] used the MK criterion to build a multi-microphone speech enhancement system without the GSC implementation and demonstrated speech enhancement with relatively little enrollment data. Applying the MK criterion to a beamformer in GSC configuration enables the beam to be steered as desired.

We demonstrate the effectiveness of our proposed techniques through a series of far-field automatic speech recognition experiments on the *Multi-Channel Wall Street Journal Audio Visual Corpus* (MC-WSJ-AV) [1]. The data was recorded in a real meeting room. It is neither artificially convoluted with measured impulse responses nor unrealistically mixed with noise recorded separately. Moreover, we investigate how much speech data is necessary to robustly estimate the active weight vectors of a GSC. We also compare the conjugate gradient algorithm with the steepest descent algorithm with the unit norm constraint for optimizing the active weight vectors.

The balance of this work is organized as follows. We review the basic concept of *independent component analysis* (ICA) and show that the pdf of subband components of clean speech is not Gaussian but super-Gaussian, and that the pdf of subband samples of speech corrupted with noise or reverberation becomes more nearly Gaussian. Section 3 reviews the definition of kurtosis. We describe our beamforming algorithm in Section 4 and then derive the objective function for estimating the active weight vectors in Section 5. In Section 6, we present the results of far-field automatic speech recognition experiments. Finally, in Section 7, we present our conclusions and plans for future work.

2. Super-Gaussian Distributions

The entire field of ICA is founded on the assumption that all signals of real interest are *not* Gaussian-distributed [6]. Briefly, their reasoning is grounded on two points:

1. The *central limit theorem* states that the pdf of the sum of

independent random variables (r.v.s) will approach Gaussian in the limit as more and more components are added, *regardless* of the pdfs of the individual components. This implies that the sum of several r.v.s will be closer to Gaussian than any of the individual components. Thus, if the original independent components comprising the sum are sought, one must look for components with pdfs that are the *least* Gaussian.

2. Entropy is the basic measure of uncertainty of information in *information theory* [6]. It is well known that a Gaussian r.v. has the highest entropy of all r.v.s with a given variance [6]. Hence, a Gaussian r.v. is, in some sense, the least predictable of all r.v.s.. Interesting signals contain structure that makes them more predictable than Gaussian r.v.s. Hence, if an interesting signal is sought, one must once more look for a signal that is *not* Gaussian.

The fact that the pdf of speech is super-Gaussian has often been reported in the literature [2, 7]. Noise, on the other hand, is typically Gaussian-distributed. In fact, the pdf of the sum of super-Gaussian r.v.s gets closer to Gaussian. Thus, a mixture signal which consists of many interference signals can be expected to be Gaussian-distributed. Based on these facts, we can remove interference signals and extract a target signal by making the pdf of the beamformer's output as super-Gaussian as possible [5].

Fig. 1 shows a histogram of real parts of subband components at f_s = 800 Hz, where we used clean speech recorded with the CTM in the SSC development set [1]. Fig. 1 also presents the likelihoods of the Gaussian and super-Gaussian univariate pdfs, the Laplace, K_0, Γ and generalized Gaussian pdfs. In Fig. 1, the parameters of the generalized Gaussian (GG) pdf are estimated from training data. As shown in Fig. 1, super-Gaussian pdfs exhibit the "spikey" and "heavy-tailed" characteristics. 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. It is clear from Fig. 1 that the distribution of clean speech is super-Gaussian.

Fig. 2 shows subband domain histograms of clean speech and speech corrupted with noise. It is clear from this figure that the pdf of the speech corrupted with noise has less probability mass around the center spike, more probability mass in intermediate regions, and less probability mass in the tail than the clean speech. This indicates 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. Kurtosis

The *excess kurtosis* or simply *kurtosis* of a r.v. Y with zero mean, defined as

$$\text{kurt}(Y) \triangleq \mathcal{E}\{Y^4\} - 3(\mathcal{E}\{Y^2\})^2, \tag{1}$$

is a measure of how *non-Gaussian* Y is [6]. The Gaussian pdf has zero kurtosis; pdfs with positive kurtosis are *super-Gaussian*; those with negative kurtosis are *sub-Gaussian*. From observed samples, we can approximate (1) as

$$\text{kurt}(Y) \approx \frac{1}{T}\sum_{t=0}^{T-1} Y_t^4 - 3\left(\frac{1}{T}\sum_{t=0}^{T-1} Y_t^2\right)^2. \tag{2}$$

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. Of the three named super-Gaussian pdfs in Fig. 1, the Γ pdf has the highest kurtosis, followed by the K_0, then the Laplace pdf; the *generalized Gaussian* (GG) pdf, on the other hand, can have arbitrarily high kurtosis depending the value assigned to the shape factor p. From Fig. 1 it is clear that as the kurtosis increases, the pdf becomes more and more spikey and heavy-tailed. The empirical kurtosis can be greatly influenced by a few samples with a low observation probability; Hyvärinen and Oja [6] note that negentropy is generally more robust in the presence of outliers than kurtosis.

4. Beamforming and Post-Filtering

Consider a subband beamformer in the GSC configuration [3, §13.6] with a post-filter. The output of a beamformer for a given subband can be expressed as

$$Y_t = (\mathbf{w}_\text{q} - \mathbf{B}\mathbf{w}_\text{a})^H \ \mathbf{X}_t, \tag{3}$$

where $\mathbf{w}_\text{q}$ is the *quiescent weight vector* for a source, $\mathbf{B}$ is the *blocking matrix*, $\mathbf{w}_\text{a}$ is the *active weight vector*, and $\mathbf{X}_t$ is the input subband *snapshot vector* at a frame t. In keeping with the GSC formalism, $\mathbf{w}_\text{q}$ is chosen to give unity gain in the desired *look direction* [3, §13.6]; i.e., to satisfy a *distortionless constraint*. The blocking matrix $\mathbf{B}$ is chosen to be orthogonal to $\mathbf{w}_\text{q}$, such that $\mathbf{B}^H \ \mathbf{w}_\text{q} = \mathbf{0}$.

This orthogonality implies that the distortionless constraint will be satisfied for any choice of $\mathbf{w}_\text{a}$. While the active weight vector $\mathbf{w}_\text{a}$ is typically chosen to minimize the variance of the beamformer's outputs, here we will develop an optimization procedure to find that $\mathbf{w}_\text{a}$ which *maximizes* kurtosis (2). In order to calculate the objective functions, the variance of the outputs Y is needed. Substituting (3) into the definition $\sigma_Y^2 = \mathcal{E}\left\{Y\,Y^*\right\}$ of variance, we find

$$\sigma_Y^2 \quad = \quad (\mathbf{w}_\text{q} - \mathbf{B}\mathbf{w}_\text{a})^H \, \mathbf{\Sigma_X} \, (\mathbf{w}_\text{q} - \mathbf{B}\mathbf{w}_\text{a})\,, \tag{4}$$

where $\mathbf{\Sigma_X} = \{\mathbf{X}\mathbf{X}^H\}$ is the covariance matrix of the input snapshot vector $\mathbf{X}$.

Maximizing the degree of super-Gaussianity yields a weight vector $\mathbf{w}_\text{a}$ capable of canceling interference including incoherent noise that leaks through the sidelobes without the signal cancellation problems encountered in conventional beamforming. Zelinski post-filtering can then be performed on the output of the beamformer [8].

For the experiments described in Section 6, subband analysis and synthesis were performed with a uniform DFT filter bank based on the modulation of a single prototype impulse response [10], which was designed to minimize each aliasing term individually.

In conventional beamforming, a *regularization* term is often applied that penalizes large active weight vectors, and thereby improves robustness by inhibiting the formation of excessively large sidelobes [3]. Such a regularization term can be applied in the present instance by defining the modified optimization criterion

$$\mathcal{J}(Y;\alpha) = J(Y) + \alpha \|\mathbf{w}_\text{a}\|^2 \tag{5}$$

for some real $\alpha > 0$, where $J(Y)$ is the empirical kurtosis. We set α = 0.1 for MK beamforming since we obtained the best recognition performance in preliminary ASR experiments.

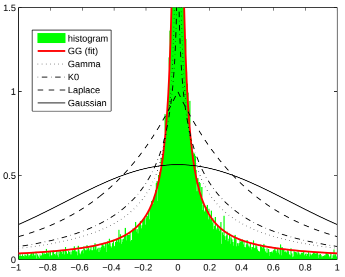

Figure 1: Histogram of real parts of subband components and the likelihood of pdfs.

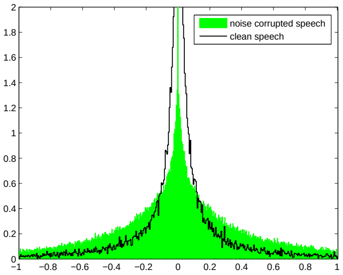

Figure 2: Histograms of clean speech and noise corrupted speech in the subband domain.

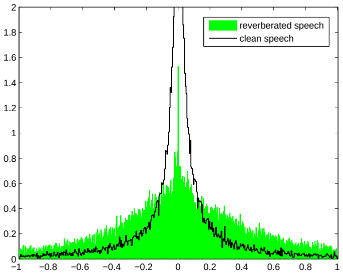

Figure 3: Histograms of clean speech and reverberated speech in the subband domain.

5. Estimation of the Active Weight Vector

With the variance of the outputs Y, σ_Y^2, the kurtosis of beamformer's output can be expressed as

$$J(Y) = \left(\frac{1}{T} \sum_{t=0}^{T-1} Y_t^4 \right) - 3 \left(\sigma_Y^2 \right)^2 . \tag{6}$$

We estimate the active weight vectors which maximize the sum of the kurtosis (6) and regularization term. In the absence of a closed-form solution, we must resort to one of the numerical optimization algorithms described below.

Upon substituting (6) into (5) and taking the partial derivative, we obtain

$$\begin{aligned} \frac{\partial \mathcal{J}(Y;\alpha)}{\partial \mathbf{w}_{\mathrm{a}}{}^*} = & \left(\frac{1}{T} \sum_{t=0}^{T-1} -2Y_t^2 \mathbf{B}^H \mathbf{X}_t Y_t^* \right) \\ & - 6\sigma_Y^2 \left(\frac{1}{T} \sum_{t=0}^{T-1} -\mathbf{B}^H \mathbf{X}_t Y_t^* \right) + \alpha \mathbf{w}_{\mathrm{a}}, \end{aligned} \tag{7}$$

Equation (7) is sufficient to implement a numerical optimization algorithm based, for example, on the method of *conjugate gradients* [11, §1.6], whereby the kurtosis of the beamformer's output can be maximized. In the experiment described in Section 6, we use the Polak-Ribiere conjugate gradient algorithm. It is compared with the steepest descent algorithm with the unit vector norm constraint.

6. Experiments

We performed far-field ASR experiments on the MC-WSJ-AV; see [1] 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 contain recordings of 10 speakers 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 the Orion source tracking system [2, 12]. Based on the average speaker position estimated for each utterance, utterance-dependent active weight vectors $\mathbf{w}_a$ were estimated for the source. The active weight vectors for each subband were initialized to zero for estimation. Iterations of the gradient algorithm were run on the entire utterance until convergence was achieved. As mentioned previously, Zelinski post-filtering [8] was performed after beamforming. 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 previous pass; see [12] for details. Through the application of cepstral mean subtraction, vocal tract length normalization, feature space adaptation, and maximum likelihood linear regression, our state-of-the-art ASR system is easily able to compensate for any frequency distortions introduced when MN or MK beamforming is performed in the subband domain.

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 with the CTM are also given in Table 1. It is clear from Table 1 that every MN beamforming algorithm provides better recognition performance than the simple delay-and-sum beamformer both without (D&S BF) and with Zelinski post-filtering (D&S BF with PF). It is also clear from Table 1 that MN beamforming with the GG pdf assumption (MN BF with GG pdf) achieves the best recognition performance. Table 1 also shows that MK beamforming (MK BF) can achieve almost the same recognition performance as MN beamforming where one utterance speech data was used for calculating active weight vectors. Notice that both MN and MK beamformers do not require speech activity detection because they are free from the signal cancellation problem seen in the minimum mean squared-error beamformer (MMSE BF) [3]. Note that the error rates given in Table 1 are to date the *lowest* reported in the literature for this ASR task [1].

In MK beamforming, the estimation of the active weight vectors is greatly influenced by outliers. We observed that the active weight vectors became extremely large in the case that the amount of data for the adaptation was insufficient. It could not be avoided even if we increased the regularization weight α. We therefore put a constraint on the active weight vector: $\|\mathbf{w}_{\mathrm{a}}\| = 1$ if $\|\mathbf{w}_{\mathrm{a}}\| \geq 1$. The active weight vector is projected on the unit circle after every step if the vector norm exceeds unity. Such a projection procedure could destroy the convergence property of the Polak-Ribiere conjugate gradient algorithm because it uses the sequence of search directions in order to approximate the curvature of the objective function around an evaluation point. Hence, we implemented the projection procedure in the steepest descent algorithm [6]. Table 2 shows the

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

Beamforming Algorithm	Pass (%WER)			
	1	2	3	4
SDM	87.0	57.1	32.8	28.0
CTM	52.9	21.5	9.8	6.7
D&S BF	80.1	39.9	21.5	17.8
D&S BF with PF	79.0	38.1	20.2	16.5
MMSE BF	78.6	35.4	18.8	14.8
MN BF with GG pdf	75.1	32.7	16.5	13.2
MK BF	76.6	33.5	17.2	13.6

Note that WERs of 12.3% for CTM and 66.5% for SDM were achieved with the adaption techniques described by Lincoln *et al* [1], who also reported that their beamforming algorithm achieved a WER of 28.1%. To the best of our knowledge, no other error rates at present have been reported in the literature on this ASR task.

Table 2: WERs for the number of frames used in adaptation for each beamforming algorithm .

Beamforming Algorithm	milli-second	Pass (%WER)			
		1	2	3	4
MN BF with Polak-Ribiere conjugate gradient	192	73.2	38.2	19.2	15.3
	384	75.7	35.0	18.9	15.4
	576	75.8	33.5	17.8	14.5
	1 utt.	75.1	32.7	16.5	13.2
MK BF with the Polak-Ribiere conjugate gradient	192	94.1	90.1	81.3	-
	384	93.3	87.2	77.0	74.7
	576	87.3	79.3	52.9	50.0
	1 utt.	76.6	33.5	17.2	13.6
MK BF with the steepest descent with the unit NC	192	80.2	41.7	21.9	18.6
	384	82.0	44.0	21.5	18.5
	576	80.1	41.1	20.5	17.5
	1 utt.	75.7	32.8	17.3	13.7

WER for the amount of data for each beamforming algorithm. It is clear from Table 2 that MN beamforming can provide good recognition performance even if very little adaptation data are available. That is mainly because the speech models trained with sufficient data are used for the calculation of negentropy. Such prior speech models make MN-beamforming robust for outliners. It is also clear from Table 2 that good recognition performance is not obtained by MK beamforming with the Polak-Ribiere conjugate gradient algorithm because the active weight vector $\mathbf{w}_a$ grows excessively large . Table 2 suggests that such a problem can be alleviated by projecting the active weight vector into the unit circle.

7. Conclusions and Future Work

In this work, we have proposed a novel beamforming algorithm based on maximizing kurtosis, which requires prior knowledge of the pdf neither of speech nor of its subband samples. We have demonstrated the effectiveness of our proposed technique through a series of large vocabulary, far-field ASR experiments on speech data captured in a *real* acoustic environment with *real* speakers; i.e., the speech material was *not* artificially convoluted with measured impulse responses, which, unfortunately, is currently the *de facto* standard way of testing beamforming and source separation algorithms. We have also investigated the relationship between the amount of data used for adaptive beamforming and recognition performance. Our results indicate that MK beamforming requires more adaptation data than MN beamforming. If, however, the amount of data is sufficient, the algorithm proposed here can achieve nearly the same performance as MN beamforming. In future, we plan to develop an on–line version of the beamforming algorithm presented here.

8. Acknowledgements

This work was supported by the European Union (EU) under the integrated projects AMIDA, *Augmented Multi-party Interaction with Distance Access*, contract number IST-033812, and by the Federal Republic of Germany under the research training network IRTG 715 ”Language Technology and Cognitive Systems”, funded by the German Research Foundation (DFG). We would like to thank Prof. Hervé Bourlard for giving us the opportunity to study about the far-field speech recognition.

9. 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. ASRU*, 2005, pp. 357–362.

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

[3] M. Wölfel and J. McDonough, *Distant Speech Recognition*. New York: John Wiley & Sons, 2008.

[4] 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.

[5] K. Kumatani, J. McDonough, D. Klakow, P. N. Garner, and W. Li, “Adaptive beamforming with a maximum negentropy criterion,” in *Proc. the Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2008.

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

[7] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, “Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 1741–1752, 2007.

[8] 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.

[9] B. W. Gillespie, H. S. Malvar, and D. A. F. Floêncio, “Speech dereverberation via maximum-kurtosis subband adaptive filtering,” in *Proc. ICASSP*, 2001.

[10] 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.

[11] D. P. Bertsekas, *Nonlinear Programming*. Belmont, Massachusetts: Athena Scientific, 1995.

[12] J. McDonough, K. Kumatani, T. Gehrig, E. Stoimenov, U. Mayer, S. Schacht, M. Woelfel, and D. Klakow, “To separate speech! a system for recognizing simultaneous speech,” in *Proc. MLMI*, 2007.