TOWARDS ONLINE MAXIMUM KURTOSIS BEAMFORMING

Kenichi Kumatani
Disney Research, Pittsburgh
4720 Forbes Avenue
Pittsburgh, PA 15213, USA
kenichi.kumatani@disneyresearch.com

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 the use of optimization criteria for beamforming that exploit the non-Gaussianity of human speech. In particular, we examined beamforming algorithms designed to maximize the kurtosis or negentropy of the sub-band output of a generalized sidelobe canceller. These techniques, while effective, require 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. In a set of distant speech recognition experiments, we compare the effectiveness of the proposed technique to several common beamformer designs. Compared to a single channel of the array, the proposed algorithm reduced word error rate from 24.0% to 10.3%, which is the best performance yet achieved on this task.

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 [1, §13.5.2]. The particular algorithms we examined were designed to either maximize kurtosis [2] or negentropy [3] of the subband output of a generalized sidelobe canceller (GSC) [1, §13]. These techniques, while effective, require 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 after a single pass through each utterance.

Distant speech recognition (DSR) has long been of interest within the automatic speech recognition (ASR) community. Recently, the effective recognition of children’s speech has become a research topic of great and growing interest within the ASR field. In this work, we combine these heretofore separate research tracks, by testing the algorithms on a corpus of far-field data. This corpus was collected with a 64-channel Mark IV microphone, lapel microphones, as well as two calibrated high definition video camcorders. The subjects of the data collection were children aged 4–10, who were either interacting with an animated character or observing cartoon-like still images on a display.

In a set of DSR experiments, we compared the effectiveness of the proposed technique to several common beamformer designs. Compared to a single channel of the array, the proposed algorithm reduced word error rate (WER) from 24.0% to 10.3%, which is the best performance yet achieved on this task. In comparison, the performance provided by a lapel microphone was 6.8%.

The balance of this work is organized as follows. In Section 2, we briefly review the non-Gaussian characteristics of speech, along with how this basic characteristic can be successfully exploited in developing effective beamforming algorithms. Also included there is a presentation of the online maximum kurtosis beamformer, the principal result of this work. The corpus used for our DSR experiments is described in Section 3. Our experimental results are tabulated in Section 4, as well as a discussion thereof. Finally, in Section 5 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 [1, §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-Gaussianity, 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

\[
\kurt(Y) \triangleq \mathbb{E}\{Y^4\} - \beta(\mathbb{E}\{Y^2\})^2,
\]

where \(\beta\) 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 [4] 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 [1, §13.6]. The output of a beamformer for a given subband at time $t$ can be expressed as

$$Y_t = (w_{q_t} - B_t w_{a_t})^H X_t,$$

where $w_{q_t}$ is the quiescent weight vector for a source, $B_t$ is the blocking matrix, $w_{a_t}$ is the active weight vector, and $X_t$ is the input subband snapshot vector. In keeping with the GSC formalism, $w_{q_t}$ is chosen to give unity gain in the desired or look direction [1, §13.6]; i.e., to satisfy a distortionless constraint. The blocking matrix $B_t$ is chosen to be orthogonal to $w_{q_t}$, such that $B_t^H w_{q_t} = 0$.

This orthogonality implies that the distortionless constraint will be satisfied for any choice of $w_{a_t}$.

While the active weight vector is typically chosen to minimize the variance of the beamformer’s outputs, here we will develop an optimization procedure to find that $w_{a_t}$ which maximizes kurtosis. Maximizing the degree of super-Gaussianity yields a weight vector $w_a$ 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 [5].

For the experiments described in Section 4, subband analysis and synthesis were performed with a uniform DFT filter bank based on the modulation of a single prototype impulse response [1, §11], which was designed to minimize each aliasing term individually.

2.3. Estimation of the Active Weight Vector

In [2], the kurtosis of the beamformer’s output was computed over an entire utterance. While such a batch algorithm is feasible for relatively small arrays of eight elements or fewer, it becomes computationally intractable for the larger array of 64 elements considered in this work. The batch algorithm also has an unacceptably slow response when a long utterance must be processed.

In this work, we calculate the kurtosis of the GSC beamformer for a block of input subband samples instead of using the entire utterance data. The kurtosis at each block $k$ can be expressed as

$$J_k(Y) = \left(\frac{1}{T_k} \sum_{t=0}^{T_k-1} |Y_t|^4\right) - \beta \left(\frac{1}{T_k} \sum_{t=0}^{T_k-1} |Y_t|^2\right)^2,$$

where $T_k$ represents a number of frames in the block.

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 [1]. Such a regularization term can be applied in the present instance by defining the modified optimization criterion

$$J_k(Y; \alpha) = J_k(Y) - \alpha \|w_{a_k}\|^2$$

for some real $\alpha > 0$. We set $\alpha = 0.1$ based on the results of the speech recognition experiments in prior work [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 (4) under the norm constraint at each block. In the absence of a closed-form solution, we resorted to the gradient descent algorithm [6, §1.6]. Upon substituting (3) into (4) and taking the partial derivative with respect to the active weight vector, we obtain

$$\frac{\partial J_k(Y; \alpha)}{\partial w_{a_k}} = -2 \left(\frac{1}{T_k} \sum_{t=0}^{T_k-1} |Y_t|^2 B_t^H X_t Y_t^*\right) + 2\beta \left(\frac{1}{T_k} \sum_{t=0}^{T_k-1} |Y_t|^2\right) B_t^H X_t Y_t^* - \alpha w_{a_k}.$$  

The gradient (5) is iteratively 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 batch method is able to track a non-stationary sound source, and provides a more accurate gradient estimate than block-by-block gradient estimation algorithms.

The beamforming algorithm can be summarized as follows:

1. Initialize the active weight with $w_{a_0} = 0$.
2. Given estimates of time delays, calculate the quiescent vector and blocking matrix.
3. For each block of input subband samples, $l = 1, 2, \cdots$, repeat estimation of the active weight vector $w_{a_l}$ based on the gradient information computed with (5) subject to the norm constraint until it converges.
4. Initialize the active weight vector for the next block and go to the step 2.

3. CORPUS

The experiments reported in the next section were run on a data set collected from July 30 to August 12, 2010 at the Carnegie Mellon Children’s School. Participants were drawn from the greater Pittsburgh metropolitan area and consisted of 28 native-English speakers, ages 4 to 10, from a range of socio-economic backgrounds. The children’s speech was captured by 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 distant speech recognition experiments, the subjects of the study were also equipped with Shure lavell microphones with a wireless connection to an RME Hammerfall Octamic II preamp and A/D converter. The Octamic II was connected via an ADAT optical cable to a RME Hammerfall HDSPe AIO sound card, which was also capable of performing the audio playback required for generating the voices of the animated characters. A PNC coaxial connection between the Mark IV and the Octamic II ensured that all audio capture and playback was sample synchronous. All capture and playback was at 41.1 kHz with a 24-bit per sample resolution.

The data collection scenario was a simple listen-and-repeat task known as Copycat, in which children were shown an illustration of an object and asked to repeat the referring phrase spoken by the experimenter (e.g., “I want the dragon’s tail,” or “Give her the crown”).

4. EXPERIMENTS

In this section, we report the results of our initial experiments with the online maximum kurtosis beamformer, and compare them to those obtained with conventional beamforming techniques.

Our basic DSR system was trained on two publicly-available corpora of children’s speech:
1. the CMU Kids’ Corpus, which contains 9.1 hours of speech from 76 speakers;
2. the Center for Speech and Language Understanding (CSLU) Kids’ Corpus, which contains 4.9 hours of speech from 174 speakers.

The data from these corpora are different in several respects.

The feature extraction used for the DSR experiments reported here was based on cepstral features estimated with a warped minimum variance distortionless response (MVDR) spectral envelope of model order 30 [1, §5.3]. Front-end analysis involved extracting 20 cepstral coefficients per frame of speech, and then performing global 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 global CMN step, then a global semi-tied covariance transform estimated with a maximum likelihood criterion [7].

HMM training was conducted along the lines suggested in [8], where several iterations of Viterbi training were followed by Gaussian splitting, followed by more iterations of Viterbi training. 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 [1, §8.1.3].

Our experiments involved three passes of speech recognition:
1. Recognize with the unadapted conventionally-trained model;
2. Estimate vocal tract length normalization (VTLN) [1, §9.1.1], maximum likelihood linear regression (MLLR) [1, §9.2.1] and constrained maximum likelihood linear regression (CMLLR) [1, §9.1.2] 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 same.

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

The search graph used in the recognition experiments reported here was constructed by initially building a finite-state automaton by stringing Copycat utterances in parallel between a start and an end state. This acceptor was convolved together with a finite-state transducer representing the phonetic transcription 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 [1, §7.3.4].

The test set consisted of four sessions of the Copycat scenario. There were a total of 354 utterances and 1,297 words spoken by the children subjects. A total of 356 utterances and 1,305 words were spoken by the experimenter.

We first tabulate the effectiveness of the speaker adaptation in the case of beamforming performed with the online maximum kurtosis algorithm, which proved to be the best, as will be subsequently shown. In order to contrast the difficulty of recognizing the speech of the children with that of the adult experimenter, we tabulate these results separately, as shown in Table 1. It is clear that the final effect of the adaptation is dramatic. The reduction in WER from the first pass to the third is four-fold and six-fold for the child and adult instructor, respectively. Note that the difference in WER between the second and third passes is no longer measurable, indicating that the remaining error is likely due to transcription errors.

Next we will investigate the effect of the filter bank on final recognition performance. To simplify matters, we will report results only for the children, and only for the final recognition pass. For the sake of convenience, these experiments were conducted with the superdirective beamformer [1, §13.3.4], which is a static design and hence less computationally expensive than the online maximum kurtosis beamformer. For computational efficiency in subband analysis, we chose a uniform DFT filter bank [1, §11.1] using the Nyquist(M) prototype design [1, §11.7]. Such a filter bank is characterized by three parameters: the number of subbands M, the total length of the prototype M × m where m is a positive integer, and the decimation factor D = M/2^r where r is a non-negative integer. These results are summarized in Table 2. It is clear from the table that increasing the number of subbands M as well as reducing the decimation factor, which is tantamount to increasing r, both lead to performance improvements. As shown in bold, the best performance was obtained with the parameter set (M, m, r) = (1024, 2, 3). Hence, these parameters will be used for the remaining set of experiments comparing beamforming techniques.

The final set of experimental comparisons is perhaps the most important given here, to wit, we directly compare several beamforming techniques holding the filter bank prototype and speech recognizer constant. Once more, we report results on the children for the third pass of recognition. These results are reported in Table 3, where for comparison the results for the single distant microphone (SDM) and lapel microphone are also shown. From the table it is apparent that the static superdirective design provides a handsome performance improvement over the SDM. The performance obtained with the simple delay-and-sum beamformer was actually worse than that from the SDM; this is perhaps due to the fact that the array was not calibrated. A large further improvement over the superdirective design was obtained, however, from the online maximum negentropy design proposed here. This result of 10.3% WER is the best result yet obtained with the far-field data of the DCAVC.
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>%WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDM</td>
<td>24.0</td>
</tr>
<tr>
<td>Delay-and-Sum</td>
<td>26.5</td>
</tr>
<tr>
<td>Superdirective</td>
<td>14.6</td>
</tr>
<tr>
<td>Maximum Kurtosis</td>
<td>10.3</td>
</tr>
<tr>
<td>Lapel Microphone</td>
<td>6.5</td>
</tr>
</tbody>
</table>

Table 3: Word error rates for three beamforming algorithms.

5. 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 both delay-and-sum and superdirective designs, both of which are static. We have also described a new corpus far-field data. This corpus was designed to support research in a great number of domains. In future, we plan to further improve the online maximum kurtosis beamformer. We also plan to develop audio-visual speaker tracking algorithms capable of both tracking the position of a speaker and simultaneously learning the mapping between microphone array and video camera coordinates.

6. REFERENCES