PARAMETER SHARING IN SUBBAND LIKELIHOOD-MAXIMIZING BEAMFORMING FOR SPEECH RECOGNITION USING MICROPHONE ARRAYS

*Michael L. Seltzer**

Speech Technology Group
Microsoft Research
Redmond, WA 98052

Richard M. Stern

Dept. of Electrical and Computer Engineering
and School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

ABSTRACT

In this paper, we present methods to improve the computational efficiency of our previously proposed algorithm for microphone array processing for speech recognition, called Subband Likelihood-Maximizing Beamforming (S-LIMABEAM). In S-LIMABEAM, the parameters of a subband filter-and-sum beamformer are optimized to maximize the likelihood of the correct transcription of the utterance, as measured by the speech recognizer itself. This approach has been shown to produce significant improvements in recognition accuracy over conventional array processing methods in a variety of noisy and reverberant environments. However, because of the manner in which recognition features are computed, the number of subband parameters that have to be jointly optimized may be large, which slows the convergence of the algorithm. To address this problem, we present two methods of sharing parameters among multiple subband filters in order to significantly reduce the number of parameters to be optimized. Both of these methods exploit the spectral smoothing that occurs in the feature extraction process, but do so in different ways. By sharing parameters in the proposed manner, we are able to obtain a significant reduction in the time to convergence of S-LIMABEAM with a minimal degradation in speech recognition accuracy.

1. INTRODUCTION

Many microphone array signal processing techniques have been proposed in the literature, *e.g.* [1]. These algorithms are almost all signal enhancement algorithms. That is, their goal is to generate an improved output *waveform*, as measured quantitatively by SNR or other distortion criteria, or qualitatively through perceptual studies. For speech recognition using microphone arrays, one of these such algorithms is used as a pre-processing step to generate an enhanced single-channel speech output signal which then gets input to the recognition engine for feature extraction and decoding.

This approach to microphone-array-based speech recognition makes the false assumption that generating an improved output waveform will necessarily result in improved speech recognition performance. However, a speech recognition system does not interpret waveform-level information directly. It is a statistical pattern classifier that operates by finding the word string that has the maximum likelihood of generating the observed sequence of feature vectors. As a result, an array processing scheme can only be expected to improve recognition accuracy if it generates a sequence of features which maximizes, or at least increases, the likelihood of the correct transcription, relative to other hypotheses.

With this in mind, we previously proposed a new approach to microphone array processing for speech recognition called LIkelihood-MAximizing BEAMforming (LIMABEAM), in which the speech recognizer itself is used to optimize the array processing parameters [2]. In LIMABEAM, a filter-and-sum beamformer is optimized to maximize the likelihood of the correct transcription, as measured by the statistical models of the speech recognizer. To specifically address speech recognition in highly reverberant environments and the problems associated with the optimization of long filters, we developed Subband LIMABEAM (S-LIMABEAM) [3], which uses a filter-and-sum architecture in the *subband* domain. Experiments showed that S-LIMABEAM was able to achieve significant improvements in speech recognition accuracy compared to conventional beamforming methods in a variety of noisy and reverberant environments. By applying subband filtering principles, the number of parameters that need to be jointly optimized is significantly reduced compared to the equivalent fullband time-domain filter-and-sum beamfomer. However, in some cases, the number of parameters that need to be jointly estimated can still be quite large, and as a result, the optimization is computationally expensive.

In this paper, we aim to improve the efficiency of S-LIMABEAM. We propose two methods of sharing parameters among different subband filters in order to reduce the total number of subband filter parameters that need to be jointly optimized. Both methods exploit the spectral smoothing that occurs in deriving speech recognition feature vectors from a frame of speech, but do so in different ways. The proposed parameter sharing techniques result in a significant reduction in the time to convergence of the S-LIMABEAM algorithm with minimal loss in performance.

The remainder of the paper is organized as follows. In Section 2, we review the S-LIMABEAM algorithm. In Section 3, we describe the two proposed methods of sharing filter parameters. In Section 4, we evaluate these parameter sharing schemes through a series of experiments. Finally, some conclusions are presented in Section 5.

2. SUBBAND LIKELIHOOD-MAXIMIZING BEAMFORMING (S-LIMABEAM)

The goal of S-LIMABEAM is to find the set of array parameters ξ that maximizes the likelihood of the correct hypothesis. We assume that the speech recognizer is an HMM-based system and that the

*This work was performed while the author was at Carnegie Mellon University.

likelihood of the correct transcription can be largely represented by the likelihood of the single most likely HMM state sequence. The log-likelihood of the correct transcription can then be expressed as

$$\mathcal{L}(\boldsymbol{\xi}) = \sum_{i=1}^{T} \log P(\boldsymbol{z}_i(\boldsymbol{\xi})|s_i) + \log P(s_1, \ldots, s_T) \qquad (1)$$

where $\boldsymbol{z}_i(\boldsymbol{\xi})$ is the feature vector for frame i, shown as a function of $\boldsymbol{\xi}$, $P(\boldsymbol{z}_i(\boldsymbol{\xi})|s_i)$ is acoustic likelihood of $\boldsymbol{z}_i(\boldsymbol{\xi})$ computed on HMM state s_i, and $P(s_1, \ldots, s_T)$ is the probability of the state sequence $\{s_1, \ldots, s_T\}$, computed from the HMM transition probabilities.

$\mathcal{L}(\boldsymbol{\xi})$ can be maximized by alternately optimizing $\boldsymbol{\xi}$ and the state sequence. For a given $\boldsymbol{\xi}$, the most likely state sequence can be determined using the Viterbi algorithm. However, the manner in which $\mathcal{L}(\boldsymbol{\xi})$ is maximized with respect to $\boldsymbol{\xi}$ is dependent on the choice of array processing parameters and feature vectors.

2.1. Subband Filter-and-Sum Processing for Speech Recognition

In this work, we exploit the processing already performed in the speech recognizer's front-end in order to perform subband filtering. Specifically, the windowing and the DFT serve as a filterbank and the downsampling is accomplished by the framing process, *e.g.* a 25-ms window and a 10-ms frame shift. Thus, we can accomplish subband processing in a manner well-suited for speech recognition applications without any additional processing.

To perform subband processing, each signal in the array is divided into a series of overlapping frames and each frame is windowed and divided into subbands via a DFT. Each subband signal from each microphone is then processed by a FIR filter. This produces a filter-and-sum architecture in the subband domain, which can be expressed as

$$Y_i[k] = \sum_{m=0}^{M-1} \sum_{p=0}^{P-1} H_p^{m*}[k] X_m^{i-p}[k] \qquad (2)$$

where $X_m^i[k]$ is the value of the DFT in subband k captured by microphone m at frame i, $H_p^m[k]$ is the pth complex tap of the subband filter assigned to microphone m and subband k and * denotes complex conjugation. M is the number of microphones and P is the length of the subband filters.

In conventional subband processing schemes, the filter coefficients $H_p^m[k]$ for a particular subband k are adapted independently from the other subbands. However, closer examination of the feature extraction process reveals that for speech recognition purposes, this is sub-optimal. In this work, we assume that mel frequency cepstral coefficients (MFCC) will be used for recognition. MFCCs are computed as the DCT of the logarithm of the mel spectrum. In turn, the mel spectrum is derived from the DFT by computing the energy in a series of weighted overlapping frequency bands. Each component of the mel spectral vector is computed as a linear combination of energy in a particular subset of DFT subbands. If we define M_i^l as the lth component of the mel spectrum of frame i and $V^l[k]$ as the value of the lth mel filter applied to subband k, this can be expressed as

$$M_i^l = \sum_{k=l_-}^{l_+} V^l[k] Y_i[k] Y_i^*[k] \qquad (3)$$

where l_- and l_+ are the DFT bins corresponding to the left and right edges of the lth mel filter, respectively. Outside of this range, the value of $V^l[k]$ is 0.

Substituting (2) into (3) clearly reveals that a given mel spectral component M_i^l is a function of the subband filter parameters of all microphones and *all subbands in the frequency range spanned by its mel filter*. Processing the subbands independently ignores this relationship. A more optimal approach would consider this set of filter coefficients *jointly* for each mel spectral component. In the next section, we describe a method of doing so.

2.2. Maximum Likelihood Estimation of Subband Parameters

In order to efficiently optimize the subband filter parameters, we assume that the components of the feature vectors are independent. This is the same assumption used by the recognizer in modeling the HMM state output distributions as Gaussians with diagonal covariance matrices. Under this assumption, the likelihood of a given state sequence can be maximized by maximizing the likelihood of each component in the feature vector independently.

We perform the parameter optimization in the log mel spectral domain, because each log mel spectral component is a function of only a subset of subbands, as shown in (3). Therefore, to maximize the likelihood of a particular vector component, we need to optimize only those subband filters required to compute that component.

We now define $\boldsymbol{\xi}_l$ to be the vector of subband filter parameters required to generate the lth log mel spectral component. $\boldsymbol{\xi}_l$ is a complex vector of length $M \cdot P \cdot (l_+ - l_- + 1)$ covering all filter taps of all microphones for the group of subbands from which the lth mel spectral component is computed. Clearly, the length of $\boldsymbol{\xi}_l$ varies depending on the width of its associated mel filter.

For each dimension of the feature vector $l = \{0, \ldots, L-1\}$, we want to maximize the log likelihood of the given HMM state sequence with respect to $\boldsymbol{\xi}_l$. Thus, if the HMM output distributions are Gaussian and the observations are log mel spectra, we perform L independent maximum likelihood optimizations of the form

$$\hat{\boldsymbol{\xi}}_l = \underset{\boldsymbol{\xi}_l}{\text{argmax}} \sum_i -\frac{1}{2} \frac{\left(\log M_i^l(\boldsymbol{\xi}_l) - \mu_i^l\right)^2}{\sigma_i^{l\,2}} \qquad (4)$$

If $\boldsymbol{\xi}_l$ represents the subband filter parameters described in Section 2.1, then $\hat{\boldsymbol{\xi}}_l$ cannot be directly optimized because the array parameters and the features are related nonlinearly. Therefore, gradient-based hill-climbing techniques must be used. The gradient vector $\nabla_{\boldsymbol{\xi}_l} \mathcal{L}(\boldsymbol{\xi}_l)$ can be computed by substituting (2) and (3) into the likelihood expression in (4) and differentiating. The gradient expression can also be similarly computed for HMM output distributions modeled as mixtures of Gaussians. A full derivation of the gradient can be found in [4].

3. SHARING SUBBAND FILTER PARAMETERS IN S-LIMABEAM

In S-LIMABEAM, L independent optimizations are performed, one for each component of the log mel spectral vector. In each optimization, $M \cdot P \cdot (l_+ - l_- + 1)$ complex parameters are jointly optimized. The values of l_- and l_+ are determined by the width of the lth mel filter. For speech recorded at 16 kHz, a 512-point DFT (corresponding to 256 subbands) and 40 mel filters are typically used. With these values, the number of DFT bins in each mel filter varies from 2 up to 23. Thus, for a modest number of microphones and a filter length of a few taps, optimizing the subband filters corresponding to the widest mel filter may require a joint optimization

of several hundred parameters. While this is significantly less than the number of parameters required for an equivalent filter-and-sum beamformer in the time domain, it is still quite high and therefore time-consuming to optimize. To alleviate this problem, we propose two methods that reduce to number of parameters that need to be jointly optimized in each of the L optimizations, and therefore make the parameter estimation more efficient.

3.1. Sharing Parameters Within Mel Spectral Components

In this approach, subband filter parameters are shared across all subbands that fall within a particular mel filter. For each log mel spectral component, a single subband filter is optimized and used by all subbands, rather than assigning a unique filter to each subband. Because each log mel spectral component is now generated using a common filter shared among all subbands, each filter parameter expressed in (2) is no longer a function of the subband index k and can be identified by solely by its microphone index m and its tap index p. As a result, a small change in the gradient computation is required.

By sharing parameters in this manner, the number of parameters is reduced from $M \cdot P \cdot (l_+ - l_- + 1)$ to simply $M \cdot P$. Clearly, the reduction in the number of parameters is proportional to the number of subbands in a particular mel filter. For the lowest frequency log mel spectral component, which is composed of only two subbands, the number of parameters is reduced by 50%. For the highest frequency log mel spectral component, composed of 23 subbands, the number of parameters required is reduced by 95.6%.

3.2. Sharing Parameters Across Mel Spectral Components

In S-LIMABEAM, the likelihood of each log mel spectral component is maximized by optimizing a set of filters applied to that component's constituent subbands. Because of each mel filter overlaps the adjacent mel filters by 50%, this results in two distinct filters for each subband, one for each of the mel components to which it contributes. This approach is well-suited to our assumption that the components of the log mel spectral vector are independent, as each component can be maximized independently, without affecting the likelihood of the other components.

Although we make this independence assumption, adjacent mel components are in fact highly correlated. The mel spectrum is derived from the energy in overlapping frequency bands such that subbands in the right half (higher frequencies) of one mel triangle are also in the left half (lower frequencies) of the next mel triangle. It is reasonable to expect, then, that for each subband, the two filters, optimized for adjacent mel components, will be similar. Therefore, we propose to reduce the number of total parameters to be estimated by optimizing a single filter for each subband, which will be used to generate both mel components.

The optimal way to estimate such a filter would be to jointly maximize the likelihood of both log mel spectral components. However, because of the overlap in subbands, jointly maximizing the likelihood of two components cannot be done without maximizing the likelihood jointly over all mel components. This requires a joint optimization over all subbands, which defeats the purpose of subband processing entirely.

Instead, we assume that for the subbands used by two mel components, the filter parameters that maximize the likelihood of one mel component also maximize the likelihood of the next component. In other words, the subband filter optimized when its corresponding subband is in the right half of one mel filter is assumed to be optimal for that same subband when it is in the left half of the next mel filter. Thus, the components of the log mel spectral vector are optimized in succession. For each component, the filters used to optimize the previous component are copied and fixed, and only the filters for new subbands (which will in turn be used for the next component) are optimized. Sharing parameters in this manner results in a 50% reduction in the number of parameters estimated, as each subband now has a single filter associated with it, rather than one for each mel component to which it contributes.

4. EXPERIMENTAL RESULTS

To test the methods of subband parameter sharing proposed in this paper, we performed speech recognition experiments on two microphone array databases. These databases were created using impulse responses recorded by a 7-element linear microphone array in rooms with reverberation times of 0.3 s and 0.47 s [5]. The inter-element spacing of the array was 5.66 cm and the speaker was directly in front of the array at a distance of 2 m. Each corpus was created by convolving utterances from the Wall Street Journal (WSJ0) test set with the appropriate set of impulse responses. We refer to each corpus with a subscript indicating the reverberation time, *i.e.* $\text{WSJ}_{0.30}$ and $\text{WSJ}_{0.47}$.

Speech recognition was performed using the Sphinx-3 speech recognition system with context-dependent continuous density HMMs (8 Gaussians/state) trained on clean speech using the WSJ0 training set which consists of 7000 utterances. The feature vectors used for recognition consisted of 13-dimensional MFCCs along with their delta and delta-delta parameters. The subband filter parameters were optimized using a parallel set of HMMs trained on log mel spectra.

In the first experiment, *Unsupervised S-LIMABEAM* was performed on the $\text{WSJ}_{0.3}$ corpus. In this method, optimization is performed based on an estimate of the true utterance transcription. For each utterance, conventional delay-and-sum beamforming was performed and the resulting features were decoded in order to estimate the transcription. Using this hypothesized transcription and the delay-and-sum features, the most likely state sequence was estimated using the Viterbi algorithm. Based on this state sequence, the subband filter parameters were then optimized and used to process the utterance. Features were extracted from the subband filter outputs and then a second pass of decoding was performed. Subband filters with only a single tap were optimized. The results are shown in Table 1.

The table shows both the Word Error Rate (WER) obtained using delay-and-sum beamforming and the proposed Unsupervised S-LIMABEAM methods as well as the relative time to convergence for the two parameter sharing methods, compared to the original S-LIMABEAM algorithm without parameter sharing. As the table shows, a 23% relative improvement over delay-and-sum processing is achieved by Unsupervised S-LIMABEAM. Furthermore, by using the proposed methods for subband filter parameter sharing we obtain a 40% reduction in the time to convergence with only a small degradation in performance. Both parameter sharing techniques still achieve an 18% relative improvement in WER over delay-and-sum beamforming.

In the second experiment, we evaluated the proposed parameter sharing methods on speech captured in an environment with substantially more reverberation, where longer subband filters are necessary for effective compensation. In this case, *Calibrated S-LIMABEAM* was performed on the $\text{WSJ}_{0.47}$ corpus. In this cor-

Array Processing Algorithm	Parameter Sharing	WER	Relative Time
Delay & Sum	-	12.8	-
Unsuper S-LIMABEAM	none	9.8	1.0
Unsuper S-LIMABEAM	share within	10.4	0.64
Unsuper S-LIMABEAM	share across	10.5	0.58

Table 1. WER and relative time to convergence for $WSJ_{0.3}$ obtained using delay-and-sum processing and Unsupervised S-LIMABEAM with and without the proposed methods of parameter sharing. Each subband filter had a single tap. The processing time is shown relative to processing without parameter sharing.

Array Processing Algorithm	Parameter Sharing	WER	Relative Time
Delay & Sum	-	59.0	-
Calib S-LIMABEAM	none	37.9	1.0
Calib S-LIMABEAM	share within	42.7	0.49
Calib S-LIMABEAM	share across	39.6	0.58

Table 2. WER and relative time to convergence for $WSJ_{0.47}$ obtained using delay-and-sum processing and Calibrated S-LIMABEAM with and without the proposed methods of parameter sharing. Each subband filter had five taps. The processing time is shown relative to processing without parameter sharing.

pus, the speech was captured in a room with a reverberation time of 0.47 s. In these experiments, the subband filter parameters were calibrated for each speaker by selecting one utterance at random to act as an enrollment utterance. For calibration purposes, the transcription of only the enrollment utterance was assumed to be known *a priori*. Using the known transcription of the enrollment utterance and features derived from the output of a delay-and-sum beamformer, the most likely state sequence was estimated as before. This state sequence was then used to optimize the subband filter parameters. In this case, each subband filter had 5 taps, effectively spanning 5 frames. After calibration, the filter parameters were fixed. No further optimization was performed and all remaining utterances for that speaker were processed using the calibrated filters. The results of this experiment are shown in Table 2.

The table shows the WER obtained using delay-and-sum beamforming, the original Calibrated S-LIMABEAM algorithm, and Calibrated S-LIMABEAM with the two methods of parameter sharing. Again, the table also shows the relative time to convergence for the proposed parameter sharing methods compared to the original S-LIMABEAM algorithm, when no parameters are shared. Using the full Calibrated S-LIMABEAM algorithm, a 36% relative improvement over delay-and-sum beamforming is obtained. By sharing subband parameters within each mel filter, the time to convergence is reduced by 51% while still achieving a 28% relative improvement over delay-and-sum processing. If the parameters are shared across mel filters, the degradation in performance is far less, and the time to convergence is still reduced by over 40%.

Comparing the results in Tables 1 and 2, it is apparent that the relative merits of the two parameter sharing methods proposed vary depending on the number of parameters, *i.e.* the length of the filters, being estimated. When only a single tap is optimized for each filter, sharing parameters within each mel filter generates better performance but sharing parameters across consecutive mel filters is more efficient. On the other hand, when the subband filter length in increased to five taps, the opposite is true. Better performance is achieved by sharing parameters across consecutive mel filters, while better efficiency is obtained from sharing parameters within each mel filter. Therefore, the decision to use one parameter sharing method rather than the other can be made by considering the desired filter length and the common trade-off between speed and accuracy.

5. CONCLUSIONS

In this paper, we have sought to improve the efficiency of the S-LIMABEAM algorithm for microphone-array-based speech recognition. In S-LIMABEAM, the parameters of a subband filter-and-sum beamformer are optimized in a manner which maximizes the likelihood of the correct transcription of the utterance, as measured by the statistical models of the recognizer itself. By optimizing the array parameters in this manner, we obtained a 29% average relative improvement in WER over conventional delay-and-sum beamfoming in environments with moderate to significant reverberation. In order to reduce the number of parameters that need to be jointly estimated in S-LIMABEAM and thus, improve the efficiency of the algorithm, we introduced two methods of sharing parameters among multiple subband filters. In one method, the parameter sharing occurs within the subbands of each mel filter, and in the other method, the sharing occurs across consecutive mel filters. By sharing filter parameters in S-LIMABEAM using the proposed approaches, we obtained an average reduction in the time to convergence of 43% and maintained a 25% relative improvement in WER over conventional delay-and-sum processing.

6. ACKNOWLEDGMENTS

This research was sponsored by the Space and Naval Warfare Systems Center, San Diego, under Grant No. N66001-99-1-8905. The content of the information in this publication does not necessarily reflect the position or the policy of the U. S. Government, and no official endorsement should be inferred. Michael L. Seltzer was supported by a Microsoft Research Graduate Fellowship.

7. REFERENCES

[1] Michael Brandstein and Darren Ward, Eds., *Microphone Arrays - Signal Processing Techniques and Applications*, Springer-Verlag, New York, 2001.

[2] M. L. Seltzer and B. Raj, "Calibration of microphone arrays for improved speech recognition," in *Proc. Eurospeech*, Aalborg, Denmark, September 2001, vol. 2, pp. 1005–1008.

[3] M. L. Seltzer and R. M. Stern, "Subband parameter optimization of microphone arrays for speech recognition in reverberant environments," in *Proc. ICASSP*, Hong Kong, April 2003.

[4] M. L. Seltzer, *Microphone Array Processing for Robust Speech Recognition*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, July 2003.

[5] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound scene database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. Int. Conf. Lang. Res. and Eval.*, Athens, Greece, June 2000.