SUBBAND PARAMETER OPTIMIZATION OF MICROPHONE ARRAYS FOR SPEECH RECOGNITION IN REVERBERANT ENVIRONMENTS

Michael L. Seltzer and Richard M. Stern
Department of Electrical and Computer Engineering and School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213 USA
{mseltzer,rms}@cs.cmu.edu

ABSTRACT

In this paper, we present a new subband microphone array processing algorithm specifically designed for speech recognition applications. We previously proposed a speech recognizer-based array processing algorithm which resulted in significant improvements in recognition accuracy when the speech was corrupted by additive noise and moderate levels of reverberation. However, little improvement was achieved over conventional beamforming methods in highly reverberant environments. Subband processing has been used to improve the poor performance of LMS-type algorithms when the number of filter parameters to estimate is large and the noise is highly correlated to the speech signal, *e.g.* in highly reverberant environments. We apply a subband approach to a new array processing architecture in which select groups of subbands are processed jointly to maximize the likelihood of the resulting speech recognition features, as measured by the recognition system itself. By incorporating the recognizer into the filter optimization scheme we ensure that signal components important for recognition are emphasized without undue emphasis on less critical components. By utilizing a subband approach, we can effectively apply this framework to highly reverberant environments. In doing so, we are able to achieve improvements in word error rate of over 20% compared to conventional methods in highly reverberant environments.

1. INTRODUCTION

Over the last few years, speech recognition systems have been deployed for a wide variety of real-world applications. In those applications where the use of a close-talking microphones is either undesirable or impractical, the use of a farfield microphone is required. As a result of the increased distance between the user and the microphone, the signal becomes more susceptible to distortion from additive noise and reverberation effects which severely degrade recognition accuracy.

In these situations, microphone arrays have used to mitigate the effects of this distortion. The corrupt speech signal is recorded over multiple spatially-separated channels which are then processed jointly to produce a cleaner output waveform. An survey of current microphone array processing methods is presented in [1]. Almost all methods proposed in the literature are speech enhancement algorithms. Their primary objective is to generate the highest quality waveform possible. As such, these algorithms are usually based on waveform-level criteria, such as signal-to-noise ratio (SNR), perceptual quality, or other such metric. In microphone-array speech recognition tasks, these methods are used as a pre-processing step to generate an improved single-channel waveform which is then passed to the recognizer for feature extraction and decoding.

This processing scheme implicitly assumes that a higher quality waveform will result in improved recognition accuracy. However, a speech recognition system does not interpret waveform-level information. Rather, it is a statistical pattern classifier whose goal is to hypothesize the correct transcription, usually accomplished by maximizing the likelihood of a set of features derived from the waveform. As a result, the 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. We believe that this is the underlying reason why many array processing methods proposed in the literature which produce very high quality output waveforms do not result in significant improvements in speech recognition accuracy compared to simpler methods such as delay-and-sum beamforming. A review of some of these methods and their recognition results was reported in [2].

With this in mind, we had previously proposed a new microphone array processing architecture designed specifically for improved speech recognition performance [3]. The array processing consisted of a filter-and-sum beamformer in which the speech recognition system itself was integrated directly into the filter design process. Unlike previous methods, the filter parameters are iteratively tuned to optimize speech recognition performance according the same maximum likelihood criterion used by the recognition engine itself.

Experiments showed that this approach can achieve significant improvements in recognition accuracy with a relatively small number of taps in environments where the speech is corrupted by additive noise and low to moderate levels of reverberation. However, our algorithm is, at its core, a gradient-descent-based LMS type of algorithm, although the objective function is significantly different from conventional adaptive filtering schemes. LMS algorithms exhibit poor convergence behavior when the noise is highly correlated to the target signal and the filter length is long, both of which are true in a reverberant environment.

In this paper, we present a new subband array processing strategy designed to improve speech recognition performance in reverberant environments. In this algorithm, the speech signal is divided into a set of independent subbands, and appropriate subsets of subbands are processed jointly to generate a maximally likely set of features for recognition. In performing array processing in this manner, we are able to achieve significant improvements in recognition accuracy in reverberant environments.

The remainder of this paper describes the proposed method and experimental results that demonstrate its effectiveness. In Section 2 we describe an array processing framework specifically designed for speech recognition. In Section 3 we apply this framework to a subband architecture designed to generate optimal recognition features. Section 4 discusses the incorporation of mean normalization of the features into the optimization framework. In Section 5, experimental results using the proposed method are shown and we present some conclusions in Section 6.

2. SPEECH RECOGNIZER-BASED MICROPHONE ARRAY OPTIMIZATION

In general, microphone array processing algorithms capture distorted speech signals $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ from *N* microphones and process them in some manner to produce an output ***z***. Typically, the array processor operates using a set of parameters Θ. Mathematically, we can express this simply as

$$\boldsymbol{Z} = f(\boldsymbol{X}, \Theta) \tag{1}$$

where ***Z*** is the information generated by the array processor.

In this work, we consider a filter-and-sum array processor whose output is processed by an HMM-based speech recognition system. We choose Θ to represent a vector of all filter coefficients of all microphones and $\boldsymbol{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_T\}$ to represent the sequence of *T* speech recognition feature vectors for which the likelihood of the correct transcription is maximum. We assume that this likelihood is largely represented by the likelihood of the most likely HMM state-sequence corresponding to the correct transcription. We can then represent the log-likelihood of the utterance as

$$Q(\boldsymbol{Z}) = \sum_{t=1}^{T} \log(P(\boldsymbol{z}_t | s_t)) + \log(P(s_1, s_2, s_3, \ldots, s_T)) \tag{2}$$

where $\boldsymbol{Z}$ represents the set of all feature vectors $\{\boldsymbol{z}_t\}$ for the utterance, *T* is the total number of feature vectors (frames) in the utterance, s_t represents state at time *t* in the most likely state sequence and $\log(P(\boldsymbol{z}_t | s_t))$ is the log likelihood of the observation vector $\boldsymbol{z}_t$ computed on the state distribution of s_t. The *a priori* log probability of the most likely state sequence, $\log(P(s_1, s_2, s_3, \ldots, s_T))$, is determined by the transition probabilities of the HMMs.

In order to maximize the likelihood of the correct transcription, $Q(\boldsymbol{Z})$ must be maximized with respect to both the array processing parameters Θ and the state sequence $s_1, s_2, s_3, \ldots, s_T$. This can be done by alternately optimizing Θ and the state sequence. For a given Θ, the most likely state sequence can be found using the Viterbi algorithm. However, because there are many layers of indirection between the array parameters the likelihood of the utterance, maximizing $Q(\boldsymbol{Z})$ with respect to Θ for a given state sequence is not as straightforward. If we assume that the state output distributions of the HMMs are modeled by single Gaussians, the log-likelihood of a given sequence of feature vectors can be expressed as

$$Q(\boldsymbol{Z}(\Theta)) = -\frac{1}{2}\sum_{t=1}^{T} (\boldsymbol{z}_t(\Theta) - \mu_{s_t})^T \boldsymbol{C}_{s_t}^{-1} (\boldsymbol{z}_t(\Theta) - \mu_{s_t}) \tag{3}$$

where the feature vectors $\{\boldsymbol{z}_t(\Theta)\}$ are now written explicitly as a function of Θ, and μ_{s_t} and C_{s_t} are the Gaussian mean vector and covariance matrix of state s_t, respectively.

In this work, we use Mel frequency cepstral coefficients as our recognition features. For a frame of speech $\boldsymbol{y}_t$, the corresponding vector of Mel frequency cepstral coefficients can be expressed as

$$\boldsymbol{z}_t = DCT(\log(\boldsymbol{W}|DFT(\boldsymbol{y}_t)|^2)) \tag{4}$$

where $\boldsymbol{W}$ represents the matrix of weighting coefficients of the Mel filters. Because of the non-linearity in (4), the log-likelihood in (3) cannot be directly maximized with respect to Θ. As a result, iterative non-linear optimization methods must be used. For example, we can compute the gradient of (3), simply as

$$\nabla_\Theta Q(Z) = -\sum_{t=1}^{T} \boldsymbol{C}_{s_t}^{-1} (\boldsymbol{z}_t(\Theta) - \mu_{s_t}) \frac{\partial}{\partial \Theta} \boldsymbol{z}_t(\Theta) \tag{5}$$

and apply hill-climbing methods to find an optimal value of Θ.

3. LOG MEL SPECTRUM SUBBAND FILTERING

Conventional FIR filter-and-sum beamforming can be expressed as

$$y(n) = \sum_{m=1}^{N} \sum_{p=0}^{P-1} h_m(p) x_m(n-p) \tag{6}$$

where *N* is the number of microphones in the array, and each signal $x_m(n)$ from microphone *m* is processed by a filter $h_m(n)$ of length *P*. In highly reverberant environments, the filter length *P* must be very large in order to compensate for the effects of reverberation. Having to jointly optimize so many parameters causes iterative gradient-based methods to exhibit poor convergence performance.

Subband filtering is a well-known approach which can alleviate these issues [4]. In traditional subband filtering, the fullband signal is bandpass filtered into a number of subbands, downsampled, and then each subband is processing independently. After processing, the subband signals are upsampled and the fullband signal is reconstructed.

In speech recognition, feature vectors are extracted from a series of overlapping frames from the utterance. Typically, a frame length of 25 ms is used, with an frame shift of 10 ms. The frame is usually then windowed and transformed into the frequency domain via a DFT, as in (4). This processing is ideally suited for a subband filtering approach, as the downsampling and the bandpass filtering are achieved without any additional computation. Furthermore, because the feature vectors are extracted from the DFT, there is no need to convert the signal back to a fullband signal. Instead, features can be extracted directly from the subbands.

Thus, the signals in each channel are segmented into a series of overlapping frames and the each frame is divided into subbands via a DFT. We then apply the same filter-and-sum approach in (6) to each individual band. However, the subband signals (*i.e.* the sequence of DFT coefficients in a particular bin) and the filter coefficients are now complex terms. Therefore, we rewrite (6) as

$$Y(t, k) = \sum_m \sum_p H_m^*(p, k) X_m(t-p, k) \tag{7}$$

where $X_m(t, k)$ is the value of the signal in the k^{th} subband captured by microphone *m* at frame *t*, and $H_m(p, k)$ is the p^{th} complex tap of the filter applied to that microphone channel and that subband and * denotes complex conjugation.

In conventional subband adaptive filtering techniques, the filter coefficients $H_m(p, k)$ for a particular subband are adapted independently from the other subbands. However, a closer examination of the feature extraction process in (4) will show that for speech recognition applications, this is sub-optimal.

For simplicity, we operate on log Mel spectra rather than cepstra. Once a sequence of log Mel spectral vectors is generated, we can generate the final sequence of cepstral vectors via the DCT. Because these two feature sets are linearly related (through the DCT), equations (3) and (5) are valid for log Mel spectra as well.

A single log Mel spectral component, $M_l(t)$ for $l = \{1, \ldots, L\}$, is computed as a weighted sum of the power spectrum computed over a fixed bandwidth. By expanding (4), this can be written as

$$M_l(t) = \sum_{k = l_1}^{l_2} W_l(k) S_Y(t, k) \tag{8}$$

where $W_l(k)$ is the value of the Mel triangle in bin *k* for the l^{th} Mel spectral component, $S_Y(t, k)$ is the power spectrum of $Y(t, k)$ defined as

$$S_Y(t, k) = Y(t, k) Y^*(t, k) \tag{9}$$

and $Y(t, k)$ is defined as in (7). The summation is computed from subband l_1 to l_2, where l_1 and l_2 are the DFT bins corresponding to the left and right edges of the l^{th} Mel triangle, respectively. The value of $W_l(k)$ outside this range is 0.

Substituting equations (7) and (9) into (8) clearly shows that a given Mel spectral component $M_l(t)$ is a function of the filter parameters of all channels *and all subbands in the frequency range of that Mel filter*. Processing each subband independently ignores this relationship. To account for this relationship, we propose to optimize this set of filter coefficients *jointly* for each Mel spectral component.

We now define Θ_l as the set of filter parameters used to generate the l^{th} Mel spectral component of the incoming speech signal. Θ_l is a complex vector of length $N \cdot P \cdot (l_2 - l_1 + 1)$ covering all channels, taps, and subbands which contribute to the l^{th} Mel component. We now derive the individual components $\partial Q(Z_l) / \partial H_m(p, k)$ of the gradient vector $\nabla_{\Theta_l} Q(Z_l)$. Note that Z_l is now a sequence of scalar values corresponding to the l^{th} log Mel spectral component of all frames.

Because we use error minimization techniques for optimization, we redefine the likelihood expression in (3) as an error function by changing its sign. Assuming that the Gaussians of the HMM state output distributions are modeled with diagonal covariance matrices, (3) becomes

$$\varepsilon_l = \frac{1}{2} \sum_t \frac{(\log(M_l(t)) - \mu_l(t))^2}{\sigma_l^2(t)} \tag{10}$$

By substituting (7), (8), and (9) into (10) and applying the chain rule, we can express the derivative of ε_l with respect to $H_m(p, k)$, evaluated for a particular values of *m, p,* and *k,* as

$$\frac{\partial \varepsilon_l}{\partial H_m(p, k)} = \sum_t \left(\frac{\log(M_l(t)) - \mu_l(t)}{\sigma_l^2(t)} \right) \left(\frac{W_l(k)}{M_l(t)} \right) \left(\frac{\partial S_Y(t, k)}{\partial H_m(p, k)} \right) \tag{11}$$

The final term in (11) can computed by substituting (7) into (9) and applying the product rule. This results in

$$\frac{\partial S_Y(k)}{\partial H_m(p, k)} = 2 X_m(t - p, k) Y^*(t, k) \tag{12}$$

Thus, using the target HMM state sequence and (10), (11), and (12), the set of filter coefficients Θ_l which will generate the maximum likelihood sequence of the l^{th} log Mel spectral components can be found using gradient-based methods. The details for one method of obtaining the target state sequence were previously described in [3].

Because the Mel filters are conventionally configured to have 50% overlap with adjacent Mel filters, each DFT subband contributes to two Mel spectral components. By processing the DFT subbands jointly for each Mel component, but independently across Mel components, the optimization of the complete log Mel spectral vector has with twice as many degrees of freedom compared to conventional subband processing. Yet, we still benefit from processing the speech in a subband basis, as we require fewer taps because of the downsampling and achieve improved convergence because each subband has a flatter spectrum than the fullband signal.

4. INCORPORATING MEAN NORMALIZATION OF THE FEATURES

Mean normalization of the features is a compensation technique in which the mean value of a sequence of feature vectors is subtracted from each of the vectors. When applied to features in the log domain, it compensates for short-term (*i.e.* less than the length of a frame) channel distortion, assuming the channel characteristics are stationary over the duration of the utterance. We would like the compensation our array processing provides to complement that produced by mean normalization. In other words, we would like the features generated by our array processing algorithm to be maximally likely for the correct transcription *after* mean normalization has been applied. To do so, it must be included in the optimization process. Because mean normalization is a linear process, it can readily be incorporated into (10) and (11).

5. EXPERIMENTAL RESULTS

To test the performance of the proposed algorithm, speech recognition experiments were performed using a reverberant microphone array corpus created from the RWCP Soundscene Database [5]. A 7-element linear array with a 5.66 cm inter-element spacing was simulated using impulse responses recorded in a tiled room with a reverberation time (RT60) of 470 ms. The user was directly in front of the array at a distance of 2 meters. To create a reverberant corpus for speech recognition experiments, utterances from the WSJ0 test set were convolved with these impulse responses to create a 7 channel reverberant WSJ0 corpus. The test set consists of 8 speakers with approximately 40 utterances per speaker.

We applied the proposed subband filter optimization scheme to the array calibration algorithm originally proposed in [3]. The calibration procedure is as follows. The user speaks an enrollment utterance with a known transcription. The captured speech is processed using delay-and-sum and 13-dimensional cepstral vectors are derived, along with their delta and acceleration coefficients. Using these cepstra and the known transcription, the optimal HMM state sequence is determined using the Viterbi algorithm. We then employ a second set of HMMs trained in a parallel manner using 40-dimensional log Mel spectra to obtain the Gaussian parameters of the output distributions of the state sequence. The subband filter parameters are then optimized for each of the 40 log Mel spectral components using conjugate-gradient descent and (10) and (11).

Once the subband filter parameters have been optimized, they are applied to all future utterances to generate log Mel spectral vectors which are then converted into cepstral coefficients and passed to the recognition system for decoding.

Speech recognition was performed using the SPHINX-III speech recognition system with context-dependent continuous HMMs (8 Gaussian/state) trained on clean speech using 7000 utterances from the WSJ0 training set.

In the first experiment, we examined the performance of the proposed algorithm as a function of the number of filter taps *P* in each subband filter. For one speaker from the test set, “440”, a calibration utterance was selected at random from the utterances in the test set at least 10 seconds long. Calibration was performed using the proposed method to generate a set of subband filters with a specified number of taps. The number of taps was varied from 1 to 8. This trains a set of filters which effectively span from 1 to 8 frames in duration. Once the filters were trained, they were applied to the remaining utterances in the test set. Figure 1 shows the word error rate (WER) as a function of filter length. For comparison, the WER achieved using conventional delay-and-sum is shown as well. As the figure shows, dramatic improvements in WER are achieved with an increased number of taps. However, the performance converges at about 6 taps. Increasing the number of taps beyond this point does not reduce the error rate and increases the risk of overfitting.

In the second experiment, the calibration was repeated for all eight speakers in the test set. For each speaker, the calibration utterance was chosen in the manner described above. Figure 2 shows the WER for all speakers for delay-and-sum processing and for the proposed filter optimization method with 3 taps per subband filter. The aggregate performance for all speakers is also shown. As the figure indicates, the calibration algorithm is able to produce relative improvements of up to 36% with an overall relative improvement of 22.6% in an environment with a 470 ms reverberation time.

6. SUMMARY

In this paper, we have presented a new subband array processing architecture designed to improve speech recognition accuracy in reverberant environments. We have proposed an optimization technique in which subsets of subband filters are optimized jointly according to the same maximum likelihood criterion used by the recognition system itself. By combining this recognizer-based optimization procedure with a subband filtering array processing architecture, we were able to overcome the limitations of our previously-proposed array processing algorithm [3] and achieve a 22.6% relative improvement in recognition accuracy over conventional methods in a highly reverberant environment.

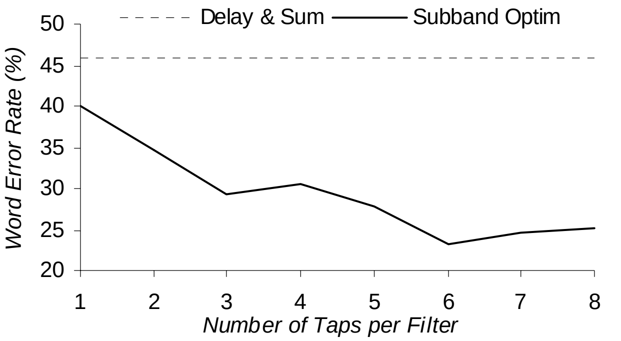

Figure 1: WER as a function of the number of taps used in the subband filters for a single speaker from the test set. The room reverberation time is 470 ms. The WER for delay-and-sum processing is 45.8%.

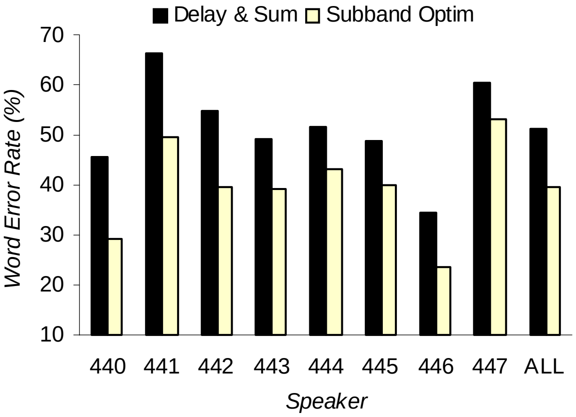

Figure 2: Word error rates for all speakers using delay-and-sum processing and the proposed calibration method with 3 taps for all subband filters. The room reverberation time is 470 ms.

ACKNOWLEDGEMENTS

The authors thank Dr. Dinei Florencio of Microsoft Research and Dr. Bhiksha Raj of the MERL Cambridge Research Lab for their suggestions regarding this work. 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 US Government, and no official endorsement should be inferred. Michael L. Seltzer is supported by a Microsoft Research Graduate Fellowship.

REFERENCES

[1] M. S. Brandstein and D. B.Ward (eds.) , *Microphone Arrays: Signal Processing Techniques and Applications,* Springer Verlag, 2001.

[2] B. W. Gillespie and L. E. Atlas, “Acoustic diversity for improved speech recognition in reverberant environments,” *Proc. ICASSP ’02,* Orlando, FL.

[3] M. L. Seltzer and B. Raj, “Calibration of microphone arrays for improved speech recognition,” *Proc. Eurospeech ’01,* Aalborg, Denmark.

[4] S. Haykin, *Adaptive Filter Theory,* Prentice Hall, NJ, 2002.

[5] http://tosa.mri.co.jp/sounddb/indexe.htm