SPEECH RECOGNIZER-BASED MICROPHONE ARRAY PROCESSING FOR ROBUST HANDS-FREE SPEECH RECOGNITION

Michael L. Seltzer[1], Bhiksha Raj[2], and Richard M. Stern[1]
1. Department of Electrical and Computer Engineering and School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213 USA
2. Mitsubishi Electric Research Labs
Cambridge, MA 02139 USA
{mseltzer,rms}@cs.cmu.edu, bhiksha@merl.com

ABSTRACT

We present a new array processing algorithm for microphone array speech recognition. Conventionally, the goal of array processing is to take distorted signals captured by the array and generate a cleaner output waveform. However, speech recognition systems operate on a set of features derived from the waveform, rather than the waveform itself. The goal of an array processor used in conjunction with a recognition system is to generate a waveform which produces a set of recognition features which maximize the likelihood for the words that are spoken, rather than to minimize the waveform distortion. We propose a new array processing algorithm which maximizes the likelihood of the recognition features. This is accomplished through the use of a new objective function which utilizes information from the recognition system itself, obtained in an unsupervised manner, to optimize the parameters of a filter-and-sum array processor. Using the proposed method, improvements in word error rate of up to 36% over conventional methods are achieved on real microphone array tasks in a wide range of environments.

1. INTRODUCTION

There are many speech recognition scenarios where the use of a close-talking microphone is either inconvenient, impractical, or unsafe. In such environments, speech signals must be captured by a microphone placed some distance away from the user. Such signals are highly susceptible to distortion from environmental noise and reverberation effects that significantly degrade speech recognition performance.

In distant-talking environments, the use of an array of microphones, rather than a single microphone, has been highly successful in compensating for this distortion. The speech signals are captured by each of the microphones simultaneously and then processed jointly using one of a variety of methods to obtain a cleaner output signal. Many microphone array processing algorithms proposed over the years have been able to achieve a substantial improvement in the quality of the output speech signal, and these methods have been used as front-ends to speech recognition systems (*e.g.* [1][2]).

Almost all array processing methods proposed to date have been developed as signal enhancement methods. When used for speech recognition, these algorithms generate the best output waveform possible, which then gets treated as a single channel input to a recognition system. The array processing component and the speech recognition system are treated as separate entities; the only communication between them is through the signal output by the array processor that is fed to the recognition system. This configuration is shown in Figure 1a. This approach to microphone-array-based speech recognition ignores a fundamental difference in the objectives of the two systems. In array processing, the goal is to produce a distortion-free waveform, as judged by SNR, human perceptual experiments or other means. On the other hand, the goal of speech recognition is to hypothesize the correct transcription of the utterance that was spoken. This usually means maximizing the likelihood of the speech recognition features derived from the waveform.

We believe that maintaining this dichotomy between the objective criteria of two individual systems limits the performance of the system as a whole. Simply put, generating an enhanced waveform does not necessarily improve recognition. The array processing scheme can only be expected to improve recognition if it results in a sequence of features which maximizes, or at least increases, the likelihood of the correct transcription. To enable this, we propose a new model for microphone array speech recognition in which the array processor and the speech recognition

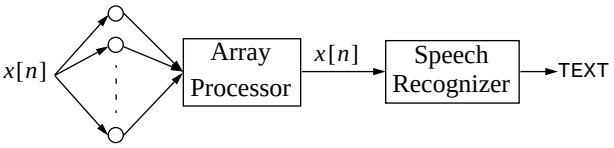

Figure 1a: Conventional microphone array speech recognition. The array processor attempts to estimate the clean waveform and the output is passed to the recognition system.

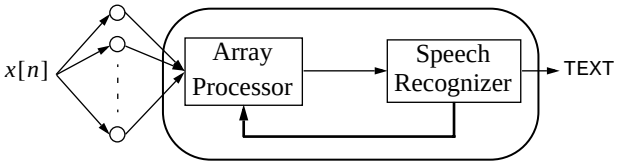

Figure 1b: The array processor and the recognition system are fully connected, allowing knowledge from the recognizer to be used in the array processing. The system no longer attempts to estimate the clean waveform.

system are treated not as two independent systems, but as two components of a single system with one common objective: to take the array input signals and generate the correct transcription. The two components of the system would be fully connected so that the vast amount of statistical information about speech contained in the recognition system could be used to guide the array processing. Such a system is shown in Figure 1b.

The main premise of this system is that an array processor can combine the signals captured by the various channels to maximize the probability that the recognition system will make a correct hypothesis. This can be achieved by choosing an array processing scheme which generates a sequence of features for which the likelihood of the correct transcription is maximum. Because likelihood is used as the criterion for recognition, we believe that optimizing the array processing to maximize the likelihood (as computed by the recognizer) of the correct transcription will increase the probability that this correct transcription will score higher than any other transcription. However, this premise results in a paradox: prior knowledge of the correct transcription is required in order to maximize its likelihood. Yet, if we had such knowledge, there would be no need for recognition!

In [4], we presented one solution to this paradox, whereby the proposed model was posed as a microphone array calibration algorithm. In this method, the user spoke an enrollment utterance with a *known* transcription. The microphone array signals and the known transcription were used in conjunction with information from the recognition system to estimate the filter parameters of a filter-and-sum array processor. The filter parameters were optimized to maximize the likelihood of the resulting recognition features for the calibration utterance. The resulting filters were then used to process all future incoming speech. The underlying assumption behind this method was that the environmental effects that degrade the calibration utterance are identical to those that degrade subsequent utterances. Thus, the optimal filters estimated for the calibration utterance would generalize to the future unknown utterances. While this method was able to achieve substantial improvements over conventional array methods on actual and simulated microphone array data, it had two drawbacks: it required the user to speak a calibration utterance, and it was unable to perform where the environmental conditions for the test utterances were different from those of the calibration utterance, *e.g.* for time-varying environments.

In this paper, we extend this initial work to optimize the array processing parameters for *each* test utterance. This approach relieves both the assumption of environmental stationarity implicit in any calibration scheme and the requirement for calibration or enrollment. Since we attempt to maximize the likelihood of the correct transcription of the test utterances, we are once again faced with the paradox of having to know the transcriptions of the very utterances that we aim to recognize. We resolve this problem by *estimating* the transcriptions, and using them in an unsupervised manner to perform the array processing.

In Section 2 we describe the proposed array processing approach. In Section 3 we show how this approach can be applied in an unsupervised manner to unknown microphone array environments. In Section 4 experimental results are shown using the proposed method, and in Section 5, some conclusions are presented.

2. SPEECH RECOGNIZER-BASED MICROPHONE ARRAY PROCESSING

In conventional feature compensation schemes for speech recognition, we search for a transformation to apply to the input speech signal that will produce a sequence of recognition features that maximizes the likelihood of the correct transcription. Within the context of a recognition system based on Hidden Markov Models (HMMs), such a transformation will take the distorted speech $x[n]$ as input and generate the sequence of feature vectors $V = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_T\}$ that is maximally likely for the HMM state sequence which generates the correct transcription.

We retain this approach to derive the optimal array processing scheme for speech recognition. If we define $X = \{x_1[n], \ldots, x_N[n]\}$ as the set of distorted speech signals captured by an array of N microphones, we look for a transformation that given X as an input, generates the most likely feature sequence for the correct transcription, V. Note that this transformation goes directly from the waveforms captured by the array to a single sequence of feature vectors, not to an output waveform.

We use Mel frequency cepstral coefficients as our recognition features and assume that the described transformation can be modeled as a filter-and-sum array processor followed by Mel frequency cepstral coefficient feature extraction. Filter-and-sum array processing can be represented as follows:

$$y[n] = \sum_{i=1}^{N} h_i[n] \otimes x_i[n - \tau_i] \tag{1}$$

where $x_i[n]$ represents the signal recorded by the i^{th} microphone, τ_i represents the delay introduced into the i^{th} channel to time-align it with the other channels, $h_i[n]$ represents the FIR filter applied to the signal captured by the i^{th} microphone, $\otimes$ is the convolution operator, and $y[n]$ represents the output signal. N is the total number of microphones in the array.

Once obtained, $y[n]$ can be parameterized to derive a sequence of feature vectors for recognition. Because the feature extraction is performed on the output of a filter-and-sum operation, the sequence of cepstral vectors can be expressed as a function of the filter coefficients of all microphone filters $h_i[k]$. If we concatenate the parameters of all filters into a supervector $\boldsymbol{h}$ and define $\boldsymbol{y}_j(\boldsymbol{h})$ as the vector of the observations for frame j expressed as a function of these filter parameters $\boldsymbol{h}$, then the vector of cepstral coefficients for frame j can be expressed as

$$\boldsymbol{z}_j(\boldsymbol{h}) = DCT(\log(\boldsymbol{M}\left|DFT(\boldsymbol{y}_j(\boldsymbol{h}))\right|^2)) \tag{2}$$

where $\boldsymbol{z}_j(\boldsymbol{h})$ represents the Mel-frequency cepstral vector for frame j and $\boldsymbol{M}$ represents the matrix of the weighting coefficients of the Mel filters. $\boldsymbol{z}_j(\boldsymbol{h})$ is, of course, a function of the filter parameters $\boldsymbol{h}$.

Now, given the sequence of maximally likely feature vectors V, and the sequence of feature vectors captured by the array, $= \{\boldsymbol{z}_1(\boldsymbol{h}), \boldsymbol{z}_2(\boldsymbol{h}), \ldots, \boldsymbol{z}_T(\boldsymbol{h})\}$, we can define an objective function which describes the error between these two feature sets as

$$= \sum_{j=1}^{T} \left\| \boldsymbol{z}_j(\boldsymbol{h}) - \boldsymbol{v}_j \right\|^2 \tag{3}$$

Of course, in a real situation, the optimal sequence of feature vectors V is unknown. In [4], we showed that if the transcription is known *a priori* (in calibration, for example), this sequence could be estimated by first using the Viterbi algorithm to estimate the most likely HMM state sequence using a sequence of feature vectors derived using a conventional array processing method, such as delay-and-sum, and then extracting the mean vectors of the state output distributions of the states in the estimated state sequence. If the HMMs have been trained from clean, undistorted speech, these means can be considered estimates of the target feature values. Because cepstral coefficients diminish in dynamic range with increasing order, the low-order cepstral coefficients will have a greater impact on the objective function than the higher order values. To prevent this, the objective function was recast in the log Mel spectral domain, where all terms have the same dynamic range. Thus the final objective function is

$$= \sum_{j=1}^{T} \left\| IDCT(\boldsymbol{z}_j(\boldsymbol{h}) - \boldsymbol{\mu}_{s_j}) \right\|^2 \quad (4)$$

where $\boldsymbol{\mu}_{s_j}$ is the mean vector of HMM state s_j. Using (1), (2), and (4), the gradient $\nabla_{\boldsymbol{h}} Q$ can be determined. Using gradient-descent-based methods, a solution to (4) can be found iteratively to obtain the optimal filter parameters $\boldsymbol{h}$.

3. UNSUPERVISED OPTIMAL FILTER-AND-SUM ARRAY PROCESSING

In the previous section, we estimated the maximally likely sequence of feature vectors $V = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_T\}$ using the HMMs of the speech recognition system and the known transcription. However, outside of a calibration scenario, the transcription is obviously unknown. We therefore estimate the transcription and use this estimate in place of the actual transcription to optimize the filter coefficients as described in the previous section.

An estimate of the transcription of the utterance can be generated by performing a first-pass recognition on the output of an initial array processing stage. This initial array processing stage can be a simple delay-and-sum, a set of filters estimated during calibration, or another array processing algorithm. Recognition features can be derived from this initial output and then decoded by the recognition system, generating a hypothesized transcription.

The hypothesized transcription can then be considered to be the actual transcription, and used with the initial set of recognition features to estimate the most likely HMM state sequence using the Viterbi algorithm. Using this state sequence, we can estimate the sequence of maximally likely feature vectors as before and optimize the filters using the objective function in (4).

Because the filter parameters are optimized using an estimate of the transcription which has been generated by the recognition system, we call this approach *unsupervised* filter optimization. Of course, the effectiveness of this algorithm is dependent on the accuracy of the hypothesized transcriptions. However, as we show in the next section, the initial transcription can have numerous errors and the algorithm can still generate a set of filter coefficients which significantly improves recognition accuracy.

corpus	# of mics	location	dist to array	jammer signal
8L	8	noisy lab	1m	none
15L	15	noisy lab	1m	none
15C1	15	conf. room	1m	none
15C3	15	conf. room	3m	none
15CR1	15	conf. room	1m	talk radio
15CR3	15	conf. room	3m	talk radio

Table 1: Description of microphone array corpora used for all experiments.

4. EXPERIMENTAL RESULTS

Experiments were performed using two real microphone array databases recorded at CMU [5], to evaluate the algorithm proposed.

The first database consists of 140 utterances (10 speakers each with 14 unique utterances), recorded using an 8 microphone horizontal linear array, with an microphone spacing of 7 cm. The array was placed on a desk in a noisy speech lab approximately 5 m x 5 m x 3 m, and the utterances were recorded with the subject seated directly in front of the array at a distance of 1 meter.

The second corpus was recorded using a 15 element log-linear array with a unit spacing of 4 cm. The corpus consists of a single male speaker in 5 different environments. An identical set of 14 utterances was recorded in each environment. This data set was recorded in both the noisy speech lab described above and a larger conference approximately 6.75 m x 5 m x 3.5 m. The conference room was quieter and more reverberant than the lab environment. The distance from the subject to the array varied between 1 m and 3 m and in some cases, a talk-radio station was playing in the background.

The utterances in both sets are comprised of alphanumeric strings and strings of command words. Each microphone array recording also has a close-talking microphone control recording for reference. Table 1 summarizes the microphone array corpora used in the experiments.

Speech recognition was performed using the SPHINX-III speech recognition system with context-dependent continuous HMMs (8 Gaussians/state) trained on clean speech using 7000 utterances from the Wall Street Journal (WSJ0) training set. Conventional 13-dimensional Mel frequency cepstral coefficients, as expressed in (2), were used as recognition features.

Each utterance was processed in the following manner. The time-delays between each of the microphone channels in the array were determined using conventional cross-correlation. The signals were then time-aligned and averaged together. This is unweighted delay-and-sum processing. Mel frequency cepstral coefficients were derived from the resulting delay-and-sum output signal and passed to the speech recognition system. A hypothesized transcription of the test utterance was generated using these features. The delay-and-sum feature stream and this estimated transcription were used to generate an estimated HMM state sequence. The mean cepstral vectors from 1 Gaussian/state HMMs correspond-

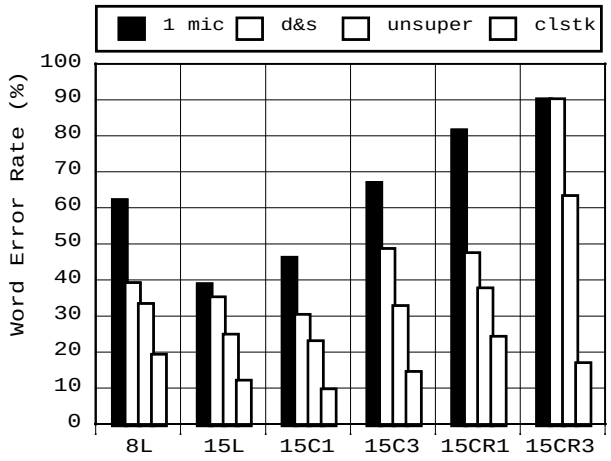

Figure 2. Word error rates for the microphone array databases using a single microphone, delay-and-sum array processing, unsupervised filter-and-sum optimization, and a close-talking microphone.

ing to the estimated state sequence were extracted The 13-dimensional cepstral vectors were converted back to 40-dimensional log Mel spectral coefficients via an IDCT, and used as estimates of the target clean speech features vectors in the array optimization for the test utterance.

Using the estimates of the clean log Mel spectra, the coefficients of FIR filters were optimized for all microphones in the array, by applying the conjugate-gradient descent method [6] to minimize (4) with respect to the filter coefficients. For expediency, the gradient-descent optimization routine was halted after 10 iterations. The resulting FIR filters were then used to process the multichannel waveforms in a conventional filter-and-sum manner. Cepstral features were extracted from this waveform and a second pass of recognition was performed. For all utterances, 50-point FIR filters were estimated for each microphone.

The recognition results for each of the databases are shown in Figure 2. As the plots indicate, the proposed speech recognizer-based unsupervised optimization algorithm is able to significantly improve recognition accuracy over delay-and-sum array processing in all environments. It is interesting to recall that the unsupervised optimization algorithm uses the hypothesized transcription from delay-and-sum processing to estimate the sequence of target feature vectors. It is quite remarkable that even when these hypothesized transcriptions are extremely poor (*e.g.* >90% WER in "15CR3"), the proposed algorithm is still able to generate filter coefficients which dramatically improve recognition. We hypothesize that this is because recognition errors are typically between acoustically confusable words, and the difference between the optimal feature sequence for the correct transcription and that for the acoustically similar words erroneously hypothesized does not adversely affect the estimation of the array filters.

It is also interesting to compare the results of the calibration method previously proposed in [4] and the unsupervised method proposed in this work. In the calibration method, a single set of optimized filters, derived from a single utterance with a *known* transcription, are applied to all future utterances. In the unsupervised case, the filters are optimized for each utterance individually, based on *hypothesized* transcriptions. In both cases, 50 point FIR filters were estimated. The performance of the two algorithms on a subset of the described databases are compared in Table 2.

Array Proc.	*8L*	*15L*	*15C1*	*15CR3*
Calibration	34.95	23.75	31.71	74.44
Unsupervised	33.53	25.00	23.17	63.42

Table 2: A comparison of word error rates for the optimal filter-and sum calibration method and the unsupervised calibration method for different microphone array databases.

The table shows that for environments where the distortion is caused predominantly by stationary noise and to a lesser extent by reverberation, as in "8L" and "15L", both methods perform equivalently. However, when the distortion is predominantly caused by reverberation or both reverberation and non-stationary noise, as in "15C1" and "15CR3" respectively, the unsupervised method is able to compensate for the distortion far more effectively.

SUMMARY

In this paper we have presented a method for microphone array processing designed specifically for speech recognition. The method optimizes the filter parameters of a filter-and-sum array processor in an unsupervised manner using information from the speech recognition system itself. By using an optimization criterion which operates explicitly in the feature domain, the algorithm is able to generate filters which emphasize signal components important for recognition, without regard to SNR or other conventional waveform-level array processing criteria. By using this algorithm, we can achieve relative improvements up to 36% over conventional delay-and-sum processing on real microphone array data in several different environments.

ACKNOWLEDGEMENTS

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.

REFERENCES

[1] T. B. Hughes, H. S. Kim, J. H. DiBiase, and H. F. Silverman, "Performance of an HMM speech recognizer using a real-time tracking microphone array as input," *IEEE Trans. on Speech and Audio Proc.*, vol. 7, pp.346-349, May 1999.

[2] P. Raghavan, R. J. Renomeron, C. Che, D. S. Yuk, and J. L. Flanagan, "Speech recognition in a reverberant environment using matched filter array (MFA) processing and linguistic tree maximum likelihood linear regression (LT-MLLR) adaptation," *Proc. ICASSP '99*, Phoenix, Arizona.

[3] D. Giuliani, M. Matassoni, M. Omologo, and P. Svaiser, "Training of HMM with filtered speech material for hands-free recognition," *Proc. ICASSP '99*, Phoenix, Arizona.

[4] M. L. Seltzer and B. Raj, "Calibration of microphone arrays for improved speech recognition," *Proc. Eurospeech '01*, Aalborg, Denmark.

[5] T. M. Sullivan, *Multi-microphone correlation-based processing for robust automatic speech recognition*, Ph.D. dissertation, Carnegie Mellon University, August, 1996.

[6] Polak, E., *Computational methods in Optimization*, New York: Academic Press, 1971.