USING A REAL-TIME, TRACKING MICROPHONE ARRAY AS INPUT TO AN HMM SPEECH RECOGNIZER

Tadd B. Hughes, Hong-Seok Kim, Joseph H. DiBiase, and Harvey F. Silverman

Laboratory for Engineering Man/Machine Systems (LEMS)
Division of Engineering, Brown University
Providence, RI 02912

ABSTRACT

A major problem for speech recognition systems is relieving the talker of the need to use a close-talking, head-mounted or a desk-stand microphone. A likely solution is the use of an array of microphones that can steer itself to the talker and can use a beamforming algorithm to overcome the reduced signal-to-noise ratio due to room acoustics. This paper reports results for a tracking, real-time microphone-array as an input to an HMM-based connected alpha-digits speech recognizer. For a talker in the very near field of the array (within a meter), performance approaches that of a close-talking microphone input device. The effects of both the noise reducing steered array and the use of a Maximum *a posteriori* (MAP) training step are shown to be significant. Here, the array system and the recognizer are described, experiments are presented, and the implications of combining these two systems discussed.

1. INTRODUCTION

Microphone-array systems have potential as a replacement for inconvenient conventional microphones, such as head-mounted, close-talking microphones or desk-stand microphones, for speech recognition applications. However, only a few attempts have been made to date to use a real-time microphone-array system for this purpose. Most early microphone-array work has been done off-line using a fixed talker with fixed beamforming[1, 2, 3] resulting in performance that has, in general, been significantly poorer than that of using a close-talking microphone. While a remote microphone system is less obtrusive, it suffers from degradation due to reverberant room acoustic artifacts and background noise. Remote microphone array systems, however, are designed to overcome these degradations. These effects are simply not significant when the talker is within a few inches of a microphone.

The novelty in this paper is the combination of a real-time, tracking microphone-array system with a modern HMM-based recognizer for the difficult English alpha-digits vocabulary. Though similar results have been shown by Yamada, Nakamura and Shikano [4, 5], the system here can be distinguished by its use of very near field, three-dimensional beamforming and an incremental-training method [6].

Two separate experiments have been performed. The first shows the recognition performance of a Baseline HMM Model incrementally retrained with microphone array data. Here, an HMM model was initially trained from a large dataset collected from talkers using a close-talking microphone for an input device. The model was then incrementally retrained with a relatively small dataset, 4500 word tokens, taken from talkers using the beamformed, microphone-array signal.

The second experiment compares, through recognition performance of the Baseline Model, the microphone array, a single remote microphone and a close-talking microphone as input devices. For this experiment an additional dataset was collected through the use of the same apparatus and testing procedure as the first, except the close-talking microphone was replaced by a single remote microphone (of the same type as the microphones used in the array) centered in the array.

The sections following briefly describe the microphone array system, the LEMS recognition system and properties of the working dataset. Next, the experiments are described and the results are presented. Finally, there is a discussion of the implications that follow.

2. THE BROWN MEGAMIKE II

A by-product of a Huge Microphone Array(HMA) project [7] is a circuit board that may be populated and packaged as a stand-alone 16-channel microphone-array system called **Brown Megamike II**. The system interfaces to a personal computer as shown in Figure 1. The microphone array itself consists of 16 pressure gradient

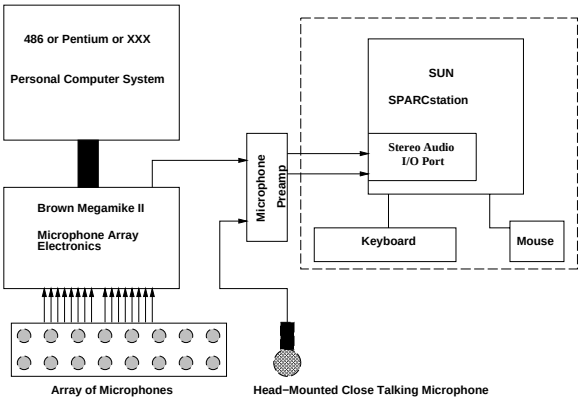

Figure 1: Block Diagram of the Apparatus

microphone elements set in rubber grommets on hardware cloth about one inch in front of a rectangular piece of acoustic foam at

This work was funded by NSF grants MIP-9120843, MIP-9314625,and MIP-9509505.

a 45 degree angle (see Figure 2) . The array is placed between the monitor and the keyboard of a SUN workstation.

All location estimation and beamforming is done by a microprocessor in the microphone-array electronics. While there is a capability for transmitting both audio and auxiliary data (e.g., pointing location) digitally, in this work the beamformed signal was passed through a DAC system and the analog signal was resampled using the recording capabilities of the SUN workstation. The

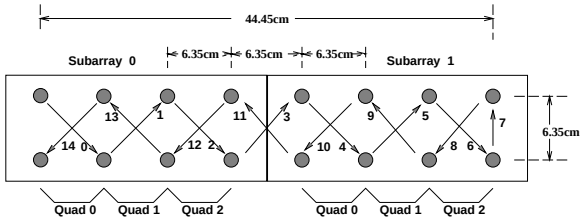

Figure 2: Microphone Array Showing Delay Estimation Pairings

processing used in these experiments is as follows:

- Signals from all 16 microphone channels are sampled simultaneously at 20ks/s to make frames of 1024 points per channel.
- The window which is applied to each frame is essentially rectangular, but has 100 points of cosine taper and 100 points of zero padding at each end. This gives 41.2ms of nonzero data.
- Sixteen, 1024-point real DFTs are taken. On the ADSP-21020 microprocessor, these are quite fast, requiring only about 365μs each.
- Fifteen time delays for the differences in the time-of-arrival of the talker signals in pairs of microphones (see Figure 2) are estimated using a weighted linear fitting and unwrapping algorithm to the phase difference in the frequency domain [8].
- Three dimensional bearing estimates are made for each *quad* (see Figure 2).
- For practical reasons, the array is considered to be two fully independent sub-arrays, each having three quads and, therefore, three spatial bearings. An estimate of the source location is made from the *close-crossing points* of the bearings of the two sub-arrays. (A *close-crossing point* is the midpoint of the line representing the shortest distance between two bearing lines.)
- When the time-delay estimates are all considered **good** (a tight cloud), then the beamformer time delays are updated for the current frame. In those cases where the estimates are not deemed **good**, the time-delay parameters for the beamformer remain unchanged.
- The array is pointed by multiplying each signal in the frequency domain by its appropriate phase shift.
- All the delayed signals are added across microphones, including the unshifted signal from the reference microphone. This is essentially unweighted *delay and sum (ds)* beamforming.
- The IDFT is taken and overlap-add is used to assemble the output.

3. THE LEMS RECOGNIZER

The LEMS recognizer, developed over the last 18 years, is an HMM-based system that has been described more fully in [9]. Its purpose is to allow the study of better signal processing and feature analysis methods for speech recognition. For this purpose, the English alphabet plus digits vocabulary is a good choice. To these 36 words are added two control words, *period* and *space*. Thus, results are gauged on a 38-word vocabulary. A tied-mixture, explicit-duration HMM is used for the recognition system. No language model is used.

Three feature sets are generated and treated as being independent; these are (1) – 12 mean-corrected low-order cepstral coefficients, (2) – their 12 first differences (in time), and (3)– a set of two features: overall energy and its first difference. The DFT-based signal processing has been optimized in earlier work [10] in which the log-magnitudes of the DFT are low-pass filtered(liftered), nonlinearly sampled, and transformed to obtain the real cepstal coefficients.

Word models are used in the LEMS recognizer. These average about five states – words like **w** have several more, and words like **e** may have as few as four states. The models are very simple as there are no *self-loops*, since explicit duration modeling eliminates this need. Skipping states is also not allowed, as experiments have shown that allowing states to be skipped generally degrades performance for the alpha-digits task with speech data obtained from talkers reading orthographic strings. Some 256 Gaussians are used for each feature space, with unique mixture coefficients for each state in each model. In training, a discrete, explicit-duration model is trained first after clustering each feature-set vector into a codebook of 256 classes. The training from the discrete system is used to initialize the training for the tied-mixture system

4. DATABASES

The largest dataset used was taken several years ago for talkers using a close-talking microphone who uttered connected strings each of about twelve alpha-digits, lasting about four seconds. One-hundred twenty talkers were recorded, each giving three sessions of 15 strings each. This totaled about 8 hours of speech data. From this, about five hours, or 3600 utterances, from 80 talkers both male and female, have been selected for training. Twenty more talkers are considered the "tweaking" set, and 20 more have been used exclusively for testing. The important statistics for this dataset are given in Table 1. All talkers have standard American dialects, with some bias due to our location in New England. This dataset is known as the Baseline dataset.

	# talkers			# utterances		
dataset	female	male	total	female	male	total
train	30	50	80	1328	2156	3484
adjust	9	11	20	243	370	620
test	8	12	20	234	361	595
total	47	73	120	1805	2887	4699

Table 1: Standard LEMS Baseline Dataset

For the experiments presented here, three new speech databases were collected from 15 male talkers and 15 female talkers who were all native speakers of American English. Table 2 gives the statistics of the new datasets. Ideally the second experiment would

dataset	male	female	# utts	# words
training (Session A)	5	5	384	4608
testing (Session A)	5	5	400	4800
testing (Session B)	5	5	362	4898

Table 2: New Simultaneously Recorded (for each Session) Datasets

require three simultaneously recorded datasets, one from the close talking headset, one from the microphone array and one from the single remote microphone. Unfortunately, the current apparatus limited concurrent recording to two channels (stereo). In order to still make a fair comparison, one set of simultaneous recordings was made with the close-talking head set and the microphone array (Session A). Recently (some eight months later), a second set of simultaneous recordings has been made (Session B). One channel took data directly from a single pressure gradient electret element mounted in the center of the array. While the other channel took the analog array output. Several of the talkers were the same as in Session A, but the room environment had changed somewhat over the interim. The new environment had more background noise in the form of "screaming" monitors and noisy disk drives. Listening to the data, one can hear a distinct improvement in the signal to noise ratio of the array signal relative to that of the single microphone.

5. INCREMENTAL MAXIMUM *A POSTERIORI* (MAP) TRAINING FOR HMM PARAMETERS

The learning technique used to adapt the recognition model is a variation on the *recursive Bayes* approach for performing sequential estimation of model parameters [11] given incremental data [12].

In the case of missing-data problems (e.g., HMMs), the Expectation-Maximization (EM) algorithm can be used to provide an iterative solution for estimation of the MAP parameters. The training algorithm that incorporates the recursive Bayes approach with the incremental EM method [12] (i.e., randomly selecting a subset of data from the training set and immediately applying the updated model) is fully described in [13] and partly in [14], thus it is not repeated here. However, it has been shown that incremental training can quickly and efficiently adapt an HMM system with a minimal amount of new data. At each update of the iterative training, there is a corresponding sequence of posterior parameters which act as the memory for previously observed data. Here, the Baseline Model is used as the initial prior information.

6. EXPERIMENTS

The large LEMS alpha-digits speech dataset described above was used to develop a Baseline HMM Model. This model became the *initial prior information* for MAP training. The testing subsets from Session A were used exclusively to generate two new recognition models: a new close-talking microphone model (CTMM), and a new microphone-array model (MAM). The new recognizers were developed by applying incremental MAP training with priors from the Baseline Model.

The objective of these experiments was to determine the impact of real-time, near-field beamforming (Sessions A and B) and of MAP training (Session A) on a complex speech recognition task. For the new data recorded in Session A, talkers properly wore a head-mounted microphone and were seated in an office-type chair that swivels and is on casters, in front of the array / workstation apparatus. They were asked to read strings from a predefined, balanced vocabulary displayed in large type on the workstation screen with no instruction given about posture or movement. However, the screen-reading task itself likely reduced movement substantially. The two recordings were not exactly simultaneous since the microphone-array speech data has some latency of about 100ms, which was estimated and corrected in every recording. The talkers for Session B were seated in the same position and used the same procedure and apparatus except for the replacement of the close-talking microphone by a single remote microphone.

Results for Session A are presented in Table 3. The performance of the Baseline Model recognizing the close-talking microphone (Session A) data is better than the same model on the Baseline dataset. Considering that the two test datasets were different, these performances are consistent. The results for the CTMM and

Test Data (Session A)	Baseline (% Correct)	CTMM (% Correct)	MAM (%Correct)
Baseline	91.7	-	-
Close Talking	94.0	94.5	-
Mic. Array	83.2	-	90.0

Table 3: Recognition Performance for Session A (% Correct)

the MAM set a new standard. The CTMM had a 94.5% performance rate on the close-talking test data (Session A), an increase of 0.5% over the Baseline Model on the same data. The MAM's 90.9% performance on the array test data (Session A) represents an increase of nearly 8% over the Baseline Model on the same data. One may conclude that incremental MAP training has significantly improved the HMM model.

The results for Session B were obtained using the Baseline Model exclusively and are given in Table 4. The raw score for the microphone array is 5.4% lower than the Baseline score in Session A. The reason for this can be found by investigating the details of how these scores were computed; the number of substitutions, insertions and deletions were summed and then divided by the total number of words spoken. It was noticed that a higher number of

Test Data (Session B)	Raw score (% Correct)	Initial Insertions Excluded (% Correct)
Single Mic.	77.2	79.0
Mic. Array	77.8	82.4

Table 4: Recognition Performance for Session B (% Correct)

insertions was being made in Session B than in Session A, while the number of substitutions roughly stayed the same. Furthermore, 57.0% of the insertions in Session B were occurring at the beginning of each utterance compared with only 11.6% for Session A. It was also noticed that there were audible artifacts at the beginning of many of the recordings made during Session B. These artifacts were due to the microphone array initially miss-aiming, possibly to a noise source, then adapting and steering correctly to the talker. In the eight month interval between Sessions A and B, background noise in the recording environment increased substantially and was

most likely responsible for the *steering problem*. When the insertions occurring *only* at the beginning of the utterances were not counted, and recordings having *any* audibly distinguishable steering artifacts were removed, the score for the microphone array increased from 77.8% to 82.4% which is consistent with Session A (83.2%). This increase in performance is good evidence that the steering problem was causing a large number of the initial insertions.

Column 1 of Table 4 shows that the raw score for the microphone array is marginally better than that for the single microphone. However, when all initial insertions were excluded [1], the microphone array out-performed the single microphone by 3.4% as shown in Column 2. As one might expect, due to the inability to reject noise, the percentage of insertions occurring at the beginning of each utterance was also high for the single microphone (38.0%), but it was even higher for the microphone array (57.0%).

7. CONCLUSIONS

When the Baseline Model was incrementally trained to the microphone array training data (Session A), the new speech recognizer, MAM, exhibited a 90.9% recognition rate on the array test data (Session A). This performance matches the highest level achieved for a close-talking microphone input in past experiments. The high quality of this performance suggests that this microphone-array system may be used as a quality input device for speech recognition systems. However, the even more recent experiment (Session B) indicates that additional work must be done to overcome the real-time steering problem in even a moderately noisy room. Future work will also involve the use of newly trained speech models for evaluating the relative performance of the single microphone and the microphone array.

An ideal input device is one which can gather speech suitable for accurate speech recognition in real-time while allowing the user freedom of movement. The current apparatus has been shown to perform accurately in a relatively quiet environment and is a step in advancing the general ease of using a modern speech recognizer.

8. REFERENCES

[1] J. Adcock, Y. Gotoh, D. J. Mashao, and H. F. Silverman. Microphone-array speech recognition via incremental MAP training. In *Proceedings of ICASSP-1996*, pages 897–900, Atlanta, GA, May 1996.

[2] A. Acero and R. M. Stern. Towards environment-independent spoken language systems. In *Proceedings DARPA Speech and Natural Language Workshop*, pages 157–162, Hidden Valley,PA, June 1990.

[3] J. L. Flanagan, R. Mammone, and G. W. Elko. Autodirective microphone systems for natural communication with speech recognizers. In *Proceedings of the Fourth DARPA Workshop on Speech and Natural Language*, pages 4.8–4.13, Asilomar, CA, February 1991.

[4] T. Yamada, S. Nakamura, and K. Shikano. Robust speech recognition with speaker localization by a microphone array. In *Proceedings of ICASSP-97*, pages –, Munich Germany, April 22-25 1997.

[5] S. Nakamura, T. Takiguchi, and K. Shikano. Noise and room acoustics distorted speech recognition by hmm composition. In *Proceedings of ICASSP-96*, pages I–69 – I–72, Atlanta, GA, May 7-10 1996.

[6] Y. Gotoh and H. F. Silverman. Incremental ml estimation of hmm parameters for efficient training. In *Proceedings of ICASSP-1996*, pages 585–588, Atlanta, GA, May 1996.

[7] H. F. Silverman, W. R. Patterson III, J. L. Flanagan, and D. Rabinkin. A digital processing system for source location and sound capture by large microphone arrays. In *Proceedings of ICASSP-1997*, pages 251 – 254, Munich,Germany, April 1997.

[8] M. S. Brandstein, J. E. Adcock, and H. F. Silverman. A practical time-delay estimator for localizing speech sources with a microphone array. *Computer, Speech and Language*, 9(2):153–169, April 1995.

[9] H. F. Silverman and Y. Gotoh. *On the Implementation and Computation of Training an HMM Recognizer having Explicit State Durations and Multiple-Feature-Set, Tied-Mixture Output Probabilities*. Number 1-1 in LEMS Monograph Series. LEMS, Division of Engineering, Brown University, Providence, RI 02912, March 1994.

[10] D. J. Mashao. *Computations and Evaluations of an Optimal Feature-Set for an HMM-based Recognizer*. PhD thesis, Brown University, 1996.

[11] M. M. Hochberg, L. T. Niles, J. T. Foote, and H. F. Silverman. Hidden Markov model/neural network training techniques for connected alphadigit speech recognition. In *Proceedings of 1991 ICASSP*, pages 109–112, Toronto, Canada, May 1991.

[12] R. M. Neal and G. E. Hinton. A new view of the EM algorithm that justifies incremental and other variants. *Biometrika*, -(-):–, – 1993.

[13] Y. Gotoh. *Incremental Algorithms and MAP Estimation: Efficient HMM Learning of Speech Signals*. PhD thesis, Brown University, 1996.

[14] Yoshihiko Gotoh, Michael M. Hochberg, Daniel J. Mashao, and Harvey F. Silverman. Incremental map estimation of hmms for efficient training and improved performance. In *Proceedings of 1995 ICASSP*, volume 1, pages 457–460, 1995.

[1] All insertions were removed from both test data sources; for the case of the single microphone only, many of these removals were potentially independent of the noise, such as a "dg" pair being recognized for "j".