Experimental Results for Baseline Speech Recognition Performance using Input Acquired from a Linear Microphone Array

Harvey F. Silverman, Stuart E. Kirtman, John E. Adcock and Paul C. Meuse

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

ABSTRACT

In this paper, baseline speech recognition performance is determined both for a single remote microphone and for a signal derived from a delay-and-sum beamformer using an eight-microphone linear array. An HMM-based, connected-speech, 38-word vocabulary (alphabet, digits, 'space', 'period'), talker-independent speech recognition system is used for testing performance. Normal performance, with no language model, i.e., raw word-level performance, is currently about 81% for a set of talkers not in the training set and about 91% for training set data. The system has been trained and tested using a close-talking head-mounted microphone. Since a meaningful comparison requires using the same speech, the existing speech database was appropriately pre-filtered, played out through a transducer (speaker) in the room environment, picked-up by the microphone array, and re-stored as a digital file. The resulting file was post-processed and used as input to the recognizer; the recognition performance indicates the effect of the input device. The baseline experiment showed that both a single remote microphone and the beamformed signal reduced performance by 12% in a room with no other talkers. For the array tested, the error is generally attributable to reverberation off the floor and ceiling.

1. Introduction

It is widely accepted that appropriate data-acquisition technology must be available in order to make speech-recognition a viable computer input mode [1, 2, 3]. While work has been done in the area of signal conditioning [4], for the last three years, research at Brown University has been in progress to develop hardware, software and algorithms as a means to make non-intrusive speech acquisition a practical reality [5, 6] Principal focus to date has been to use the phase relationships among a group of microphones spaced in a line - hence a *linear* array - for the remote, real-time acquisition of a talker's data. Various beamforming and talker location/tracking algorithms have been studied, reported, and evaluated relative to listening quality [7, 8, 9, 10, 11, 12]

The quality of a speech data acquisition system may be assessed in several ways. For many applications, evaluation is usually given, quantitatively, in terms of some signal-to-noise measure or human-listening experiment score, or qualitatively in terms of human evaluation. However, for a system whose output is fed to a speech recognizer, the recognition performance is an excellent, quantifiable measure; this approach and its results make up the body of this paper.

A key problem for such systems to overcome is that of reverberation. Acoustic reflections in a normal room environment make the output of a remote microphone quite different from that taken from the normal, close-talking, recognizer microphone. Several ways have been suggested to alleviate this problem:

- A more *focused* array system will attenuate reflections coming from a wider off-axis volume[13]. Many microphones are required to do this, and a system with beam-width control over a broad spectrum and in two or three directions is essential. This is the *spatial-filtering* approach to solving the problem.
- The acoustic environment near the microphones is very critical. New ways of mounting the microphones in an appropriately sound absorbent material substantially improve performance, without necessarily limiting the practicality of the array. More directional elements can also be used. This is an *acoustical* approach to helping to resolve the problem.
- One form or another of deconvolution can be used to undo the effects of reverberations [3, 14, 15, 16, 17, 18, 19]. Either directly or indirectly, some characterization of the room is obtained, usually as some spatially-dependent impulse response. After this non-trivial problem is solved, some processing "art" is often essential to overcome nulls in the spectrum and perform inverse filtering.

This project investigates all of the above methods. It might be added that, when working with real acoustic systems, mechanisms for reducing reverberations must be carefully applied; it is a hard problem. However, the purpose of this paper is not to deal with the improvements achieved by employing various means to dereverberate the output signal of the array; rather, it is to set a baseline standard against which to compare future developments. The problem is posed: how badly does recognizer performance degrade when the input signal is from 1) a single remote omnidirectional microphone,

*Supported principally by NSF/DARPA Grant No. IRI-8901882

cr from 2) the beamformed output from a linear microphone array? This experiment quantifies the acceptability (or lack thereof) of using relatively straightforward implementations of remote microphone technology for speech recognition.

2. The LEMS Speech Recognizer

An HMM-based, connected-speech, 38-word vocabulary (alphabet, digits, 'space', 'period'), talker-independent speech recognition system has been running for two years in the LEMS facility [20, 21]. This small, but very difficult vocabulary has many of the problems associated with a phoneme recognizer.

Speech, sampled at 16kHz from a close-talking microphone, is truncated through a 40ms Hamming window every 10ms. Twelve cepstral coefficients, twelve delta cepstral coefficients, overall energy and delta overall energy comprise the 26 element feature vector. Three 256-entry codebooks are used to vector quantize the data from cepstral, delta cepstral, and energy/delta energy features respectively [1]. The recognizer differs from standard HMM models in that durational probabilities are handled explicitly [22]. For each state, self transitions are disallowed. During training, nonparametric statistics are gathered for each of 30 potential durations in the state, i.e., 10ms to 300ms. In the base system used for this experiment, a gamma distribution was fitted to the nonparametric statistics. The models used are word-level models having from five to twelve states. Only forward transitions and skips of a single state were allowed.

The best available recognizer at the time was used for the experiment, except that the amount of speech normally used to develop the vector quantization codebooks was reduced from one and one-half hours to 15 minutes. This made it feasible to do several full k-means re-trainings of the system; VQ training took but two days (elapsed time) on a SUN SPARCstation 2 while VQ training for the one and one-half hour case would have taken an unacceptable twelve days[2]! The change to the VQ training degraded performance for the close-talking microphone data by 1.5%, i.e., the 79% performance of the system for 1) new talkers and 2) no grammar was reduced to 77.5%.

About four hours of speech (2400 connected strings, or nearly 40,000 vocabulary items) from 80 talkers, half male, half female, were used to train the hidden Markov models. Currently, the training procedure requires 60 hours of CPU time from each of eight SPARC 1+/2 workstations linked in a loosely-coupled fashion through sockets. Well-known mechanisms for speeding up the process, such as doing the computation in the logarithm domain using integers and a lookup table [23], as well as some detailed new programming speedups [24] are being used to reduce the training time.

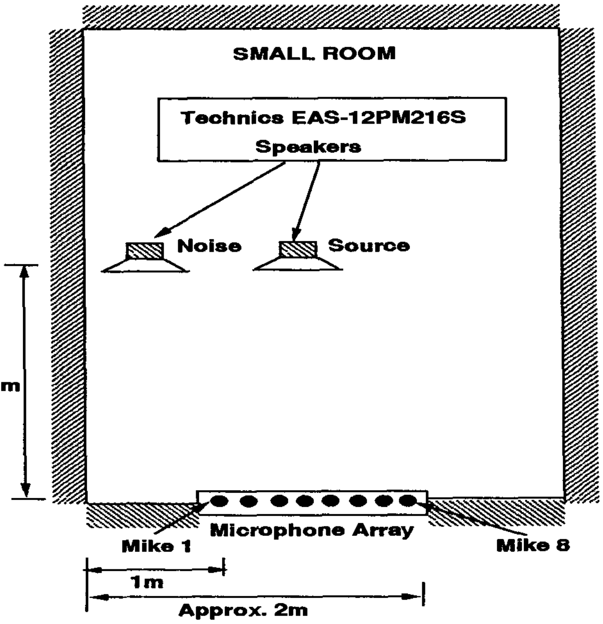

Figure 1: Acoustical Geometry of Array/Sources

3. Data Development

The original speech data were recorded in a large, generally not-too-noisy room through an Audio Technica ATM73a head-mounted, close-talking microphone. The speech was sampled through a Sony DAT- 16 bits at 48kHz sampling rate. It was then digitally decimated to 16kHz and fed directly to a SUN workstation to build a high-fidelity database [25]. The signal-to-noise ratio is about 50dB.

It would not have been possible, let alone feasible, to record another large dataset from the same talkers using the microphone array system for acquisition. Thus, a mechanism had to be developed to use the high-fidelity database as input to the array recording system. A high-quality transducer was used to play out the speech; the geometry is shown in Figure 1. The resulting real-time system for the data conversion is schematically shown in Figure 2. Three SPARC 1+/2 workstations are used. The first converts the digital speech data in *speech recognition format* into digital data acceptable for playback through the microphone array hardware. This involves changing the sampling rate from 16kHz to 20kHz and then applying an FIR *inverse filter* to undo the coloring that will come from the output transducer. This filter was obtained by running digital, band-limited white noise with DFT spectrum $W(r)$ through the transducer and recording the output through an ultra-flat frequency response Brüel & Kjær (B&K) condenser microphone system placed a few

[1] At the time this experiment was initiated, semi-continuous modeling of output probabilities and better word models were not yet a part of the system. Current improvements have increased overall performance for the head-mounted microphone input by about 3%.

[2] We are optimizing this program now.

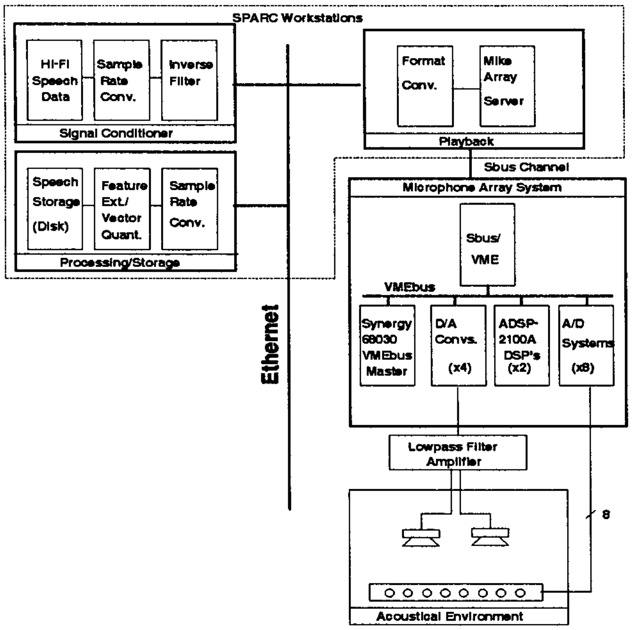

Figure 2: The Data Conversion System

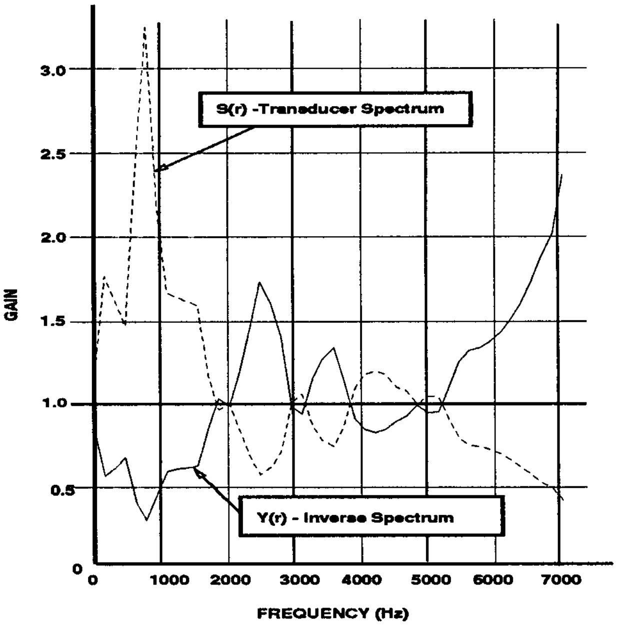

Figure 3: Spectra of the Output Transducer System and Inverse Filter

centimeters in front of the middle of the output transducer. After accumulating an average magnitude spectrum of the B&K's output via multiple 128-point DFT's, the spectrum $S(r)$ was inverted, i.e., $Y(r) = W(r)/S(r)$, and inverse transformed to produce a zero-phase FIR filter[3]. Any spectral energy attenuated by the anti-aliasing filter i.e., frequencies above 7kHz, were forced to unity gain. $S(r)$ and $Y(r)$ are shown in Figure 3. The subjective audible effects as well as the flattened white-noise response indicate that this procedure was successful in removing the 'boominess' potentially introduced by the transducer system.

Initially, small, omnidirectional electret microphones were mounted at the edge of a 5cm×10cm board containing amplifier/filter electronics and the board was plugged vertically into a (2.5m) horizontal cage. Recent work disclosed that this system formed resonant cavities that impacted the performance of the linear microphone array. When the same microphones with the same spacing (18.4 cm) were inserted into a (180cm×30cm×15cm) block of six pound ester foam, the degradations due to the cavities disappeared as may be seen in Figure 4. Note that the data shown are for the transducer output after the noise has been inverse filtered.

The remainder of the data conversion system is straightforward. Twenty kilohertz sampling interrupts are used both to produce the speech output(s) and to digitize the analog signals from the eight microphones. Sufficient memory is available for about 10 second utterances. Upon completion of an utterance, the microphone data are sent to a third SPARCstation for sample-rate conversion, signal processing for recognition, and archiving on hard disk as feature vectors for the recognition system.

4. Experiment and Results

The system was trained, both for VQ and for the hidden Markov model parameters, three different times: 1) for the high-fidelity data, 2) for the output of a single microphone of the array (a central one), and 3) for the simple delay-and-sum beamformed output of the 8 microphone array. The recognizer was tested using 20 new talkers, again half male and half female, for a total of an hour of speech, or about 4800 vocabulary items. The data conversion system was run under 'quiet' conditions. Not including noise due to reverberations, the signal-to-noise ratios were significantly degraded by the acoustical noise to 24dB for the single remote microphone and 26dB for the beamformed signal. The results as a function of talker number are plotted in Figure 5. From the Figure, one may deduce that:

- For all cases, variation with respect to talker is far greater than variations due to other effects.
- Recognition performance is approximately the same for the single microphone as it is for the beamformed case, given no other point 'noise' sources..
- Performance for the high-fidelity signal is consistently about 12% better than for the acoustically degraded signal.

[3] Non-zero-phase inverse filters are also being investigated.

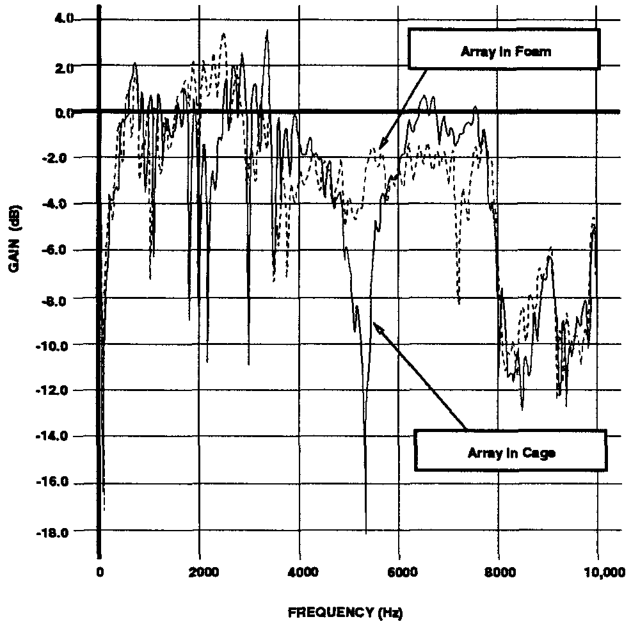

Figure 4: Spectra of Array Microphone Response in Cage and Foam

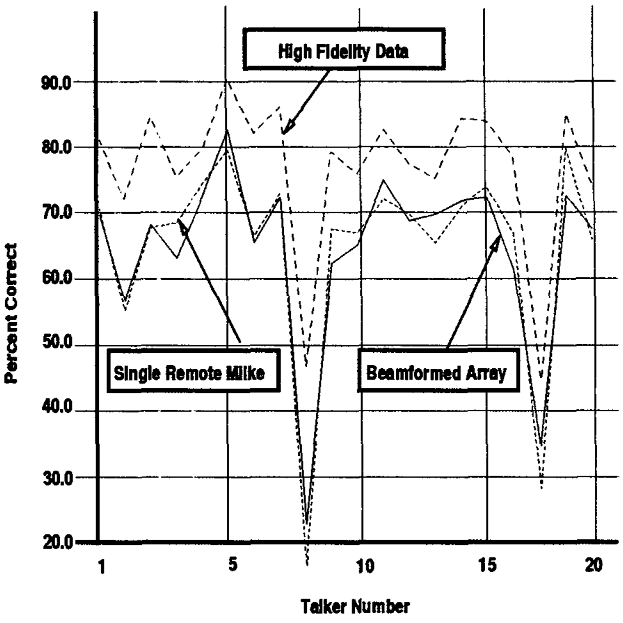

Figure 5: Recognition Performance for the Three Acquisition Systems

For completeness, each of the test datasets was run against each of the three systems. The results are given in Table 1.

Test Data	Model Trained from		
from	Hi-Fi	Remote Mike	Beamformed
Hi-Fi	77.5%	50.0%	53.6%
Remote Mike	38.8%	65.6%	64.6%
Beamformed	32.8%	57.6%	65.3%

Table 1

Averaged Results for Direct and Cross-Trained Systems

5. Discussion

Given the degraded acoustical environment, it was not surprising that performance for the converted data was reduced using remote-microphone input. However, it was somewhat surprising that this very carefully done experiment indicates no performance advantage when simple beamforming is used to generate the input. This could be due to the following:

- Low-frequency background noise is not effectively eliminated by an acoustic array of this type and size. Some filtering, perhaps combined with sub-band types of enhancements, should help.
- The major reverberations in the room come from the ceiling and floor. They have been measured as being as much as 25% of the original wavefront in intensity. Even if the reflections average 10%, implying a 14dB signal-to-noise ratio, 'quiet' room conditions no longer hold. A focused two or three dimensional array could attenuate these reflections and thus address the problem. Alternatively, pressure gradient microphones could be used in a one-dimensional array as done in [13].
- There is always some variability in an acoustical experiment regarding equipment positioning, overall amplitudes, microphone calibration etc. While great care was taken, certainly the beamformer output would be more susceptible to these variabilities than would be the single remote microphone.

In order to determine the impact of beamforming, the testing data were run through the data conversion system (source at $(1m, 2m)$) several additional times, each with a second transducer located at $(2m, 2m)$. This second transducer repeated a few seconds of speech at various, controlled levels as the testing data were being recorded. This procedure permits the assessment of the effects of beamforming with respect to spatial filtering of off-axis noise. The test datasets for both the single remote microphone and the beamformed data were run through their respective quiet-room recognizers. As the purpose of this test was to check the simple beamformer, more elaborate beamformers were not used to generate the data of Figure 6. Also, note that no background noise processing (such as high-pass filtering the signals) was used to remove the low-frequency 'rumble' of the room.

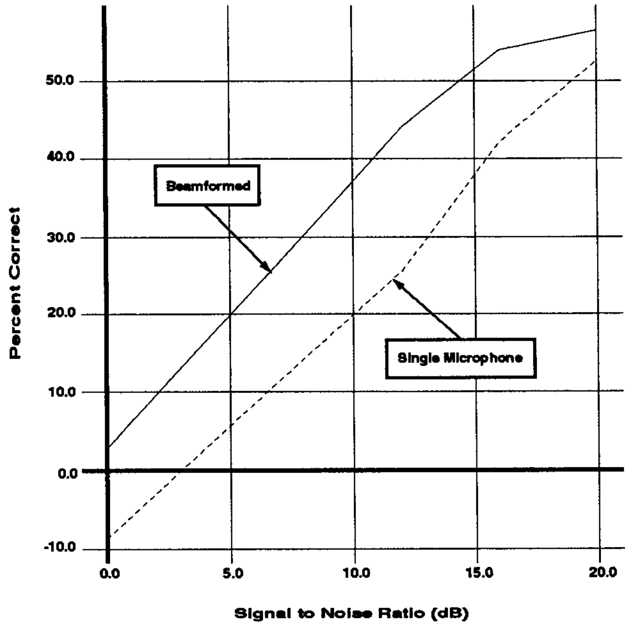

Figure 6: Performance in a Noisy Environment

As the graph indicates, there is an appreciable performance gain using the array for acoustic data collection in a noisy environment. The simple beamformer consistently scored 10-15% higher than a single microphone for SNR's less than 16dB. Note that in one case the recognition result is negative. This is a consequence of the method employed for calculating the performance score.

6. Conclusion

An intricate experiment has been developed to quantify the effects of alternative acoustic environments on speech-recognition systems. The performance of an HMM-based alphadigit recognizer was reduced about 12% when input was converted from high-fidelity, close-talking input to either a single remote microphone or the output of a delay-and-sum beamformer using an eight-microphone linear array under quiet conditions. Beamforming did significantly improve performance over that of a single microphone for low signal-to-noise ratios and is thus advantageous in the presence of acoustic interference.

More importantly, though, the work establishes an automated procedure for reconstructing a given database in a new environment, permitting the evaluation of acoustic-input devices. Such a structured methodology has allowed the determination of baseline performance and now future improvements can be appropriately measured.

References

[1] J. L. Flanagan. Bandwidth design for speech-seeking microphone arrays. *Proceedings of 1985 ICASSP*, pages 732-735, March 1985.

[2] J. L. Flanagan, J. D. Johnson, R. Zahn, and G. W. Elko. Computer-steered microphone arrays for sound transduction in large rooms. *J. Accoust. Soc. Am.*, 78(5):1508-1518, November 1985.

[3] J. B. Allen, D. A. Berkley, and J. Blauert. Multimicrophone signal-processing technique to remove room reverberation from speech signals. *Journal of the Acoustical Society of America*, 62(4):912-915, October 1977.

[4] A. Acero and R. M. Stern. Towards environment-independent spoken language systems. In *Proceedings of the Speech and Natural Language Workshop*, pages 157 - 162, Hidden Valley, Pennsylvania, June 1990.

[5] H. F. Silverman. Some analysis of microphone arrays for speech data acquisition. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, ASSP-35(2):1699-1712, December 1987.

[6] M. S. Brandstein. Design and implementation of the LEMS microphone array for the acquisition of high quality speech signals. Brown University Honor's Thesis, May 1988.

[7] V. M. Alvarado and H. F. Silverman. Experimental results showing the effects of optimal spacing between elements of a linear microphone array. In *Proceedings of 1990 ICASSP*, pages 837-840, Albuquerque, NM, April 1990.

[8] V. M. Alvarado. *Talker Localization and Optimal Placement of Microphones for a Linear MIcrophone Array using Stochastic Region Contraction*. PhD thesis, Brown University, May 1990.

[9] H. F. Silverman. DSP beamforming and talker tracking with a linear microphone array. In *Proceedings of JASA 119th Meeting*, page S3, State College, PA, May 1990.

[10] H. F. Silverman. An algorithm for determining talker location using a linear microphone array and optimal hyperbolic fit. In *Proceedings DARPA Speech and Natural Language Workshop*, pages 151-156, Hidden Valley,PA, June 1990.

[11] M. Berger and H. F. Silverman. Microphone array optimization by stochastic region contraction (SRC). *IEEE Transactions on Signal Processing*, 39(11):2377-2386, 1991.

[12] H. F. Silverman and S. E. Kirtman. A two-stage algorithm for determining talker location from linear microphone-array data. LEMS Technical Report 97, LEMS,Division of Engineering, Brown University, Providence, RI 02912, November 1991.

[13] 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.

[14] S. T. Neely and J. B. Allen. Invertability of a room impulse response. *Journal of the Acoustical Society of America*, 66(1):165-169, July 1979.

[15] R. Zelinski. A microphone array with adaptive postfiltering for noise reduction in reverberent rooms. In *Proceedings of ICASSP88*, pages 2578-2580, New York, NY, April 1988.

[16] D. Van Compernolle, W. Ma, F. Xie, and M. Van Diest. Speech recognition in noisy environments with the aid of microphone arrays. *Speech Communication*, 9(5/6):433-442, December 1990.

[17] D. Bees, M. Blostein, and P. Kabal. Reverberent speech enhancement using cepstral processing. In *Proceedings of ICASSP91*, pages 977- 980, Toronto, Ont, Canada, April 1991.

[18] H. Wang and F. Itakura. An approach of dereverberation using multi-microphone sub-band envelope estimation. In *Proceedings of ICASSP91*, pages 953- 956, Toronto, Ont, Canada, April 1991.

[19] H. Yamada, H. Wang, and F. Itakura. Recovering of broad band reverberent speech signal by sub-band mint method. In *Proceedings of ICASSP91*, pages 969- 972, Toronto, Ont, Canada, April 1991.

[20] L. T. Niles and H. F. Silverman. Combining hidden Markov model and neural network classifiers. In *Proceedings of 1990 ICASSP*, pages 417-420, Albuquerque, NM, April 1990.

[21] M. M. Hochberg, J. T. Foote, and H. F. Silverman. The LEMS talker-independent speech recognition system. LEMS Technical Report 82, LEMS, Division of Engineering, Brown University, Providence, RI 02912, March 1991.

[22] M. M. Hochberg, J. T. Foote, and H. F. Silverman. Explicit state duration modeling for HMM-based connected speech recognition. In *1991 Arden House Workshop on Speech Recognition*, Harriman, NY, December 1992. Accepted for Presentation.

[23] P. F. Brown. *The Acoustic Modeling Problem in Automatic Speech Recognition.* PhD thesis, Carnegie-Mellon University, Pittsburgh, PA, May 1987.

[24] J. T. Foote, M. M. Hochberg, P. M. Athanas, A. T. Smith, M. E. Waslowski, and H. F. Silverman. Distributed hidden Markov model training on loosely-coupled multiprocessor networks. In *Proceedings of 1992 ICASSP(Accepted)*, San Francisco, CA, March 1992.

[25] J. T. Foote. The LEMS DAT/SCSI interface: User's guide and technical reference. LEMS Technical Report 95, LEMS, Division of Engineering, Brown University, Providence, RI 02912, November 1991.