COMPARISON OF AUDITORY MODELS FOR ROBUST SPEECH RECOGNITION*

Charles R. Jankowski Jr. and Richard P. Lippmann

MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02173

ABSTRACT

Two auditory front ends which emulate some aspects of the human auditory system were compared using a high performance isolated word Hidden Markov Model (HMM) speech recognizer. In these initial studies, auditory models from Seneff [2] and Ghitza [4] were compared using both clean speech and speech corrupted by speech-like "babble" noise. Preliminary results indicate that the auditory models reduce the error rate slightly, especially at intermediate and high noise levels.

1. MOTIVATION

The performance of speech recognizers often degrades dramatically in noise, with different talking styles, when the microphone is changed, and as a talker moves relative to a microphone. New auditory front ends that mimic some aspects of human auditory-nerve and psychoacoustic behavior have been proposed to reduce these problems. Although past limited experiments suggest that these front ends improve robustness, no thorough comparisons have been performed using high-performance Hidden Markov Model (HMM) recognizers. In addition, few studies have evaluated the effect of speech babble noise and frequency response variability on performance or explored alternative approaches to feature reduction. This is an early progress report on research in this area. Further ongoing experiments are exploring additional front ends and alternative data reduction techniques.

2. AUDITORY MODELS

Two auditory front ends which produce features that correspond to phase or synchrony information in the speech signal were explored. These auditory front ends were compared to a more conventional mel-scale cepstral front end.

*This work was sponsored by the Defense Advanced Research Projects Agency. The views expressed are those of the authors and do not reflect the official policy or position of the U. S. Government.

2.1. Mel-Scale Cepstra

The mel-scale cepstral front end described by Davis and Mermelstein [1] was used as a reference. This is a common signal representation which is currently used in all speech recognition systems at Lincoln Laboratory. In this front end, a 20 ms Hamming window is applied to the speech signal every 10 ms. The power spectrum of the windowed waveform is weighted by a series of filters, linearly spaced from 0 to 1000 Hz, and logarithmically spaced above 1000 Hz. Each filter "width" is twice the spacing. An inverse cosine transform converts the logarithm of the resulting filter bank coefficients to the cepstral domain.

2.2. Seneff's Auditory Model

The first auditory front end evaluated was motivated from physiological data and described by Seneff [2,3]. This front end incorporates a first stage of 40 linear filters, followed by a series of nonlinearities modelling the transformation from basilar membrane motion to auditory nerve stimulation. Such nonlinearities include soft half-wave rectification, a model for short-term adaptation, and a rapid AGC.

Seneff's front end had two outputs. "Mean rate" outputs are generated by detecting the envelope of the nonlinear stage output. These outputs roughly corresponds to spectral magnitude information. "Synchrony" outputs detect the extent that the nonlinear stage output for a particular channel has energy at the center frequency for that channel. This emulates the extent to which the nerve firings from a particular location of the basilar membrane are synchronized to the "characteristic frequency" corresponding to that location.

2.3. EIH

The second auditory front end evaluated was the Ensemble Interval Histogram (EIH) model developed by Ghitza [4]. The EIH model has a first stage of linear filters similar to Seneff's, but with a considerably higher number of filters (133 instead of 40). The second stage takes the output of each filter and computes the intervals between the positive crossings of the filtered waveform at various logarithmi-

cally spaced thresholds. A histogram of the frequencies corresponding to these intervals is then created. The final stage combines the histograms for each of the channels together into the final output, the Ensemble Interval Histogram. In this respect, the EIH model performs functions similar to Seneff's "Synchrony" output; measuring the extent that the output of the linear filter is in synchrony with the center frequency of that filter. The EIH model has been shown useful in performing isolated-word recognition in high noise conditions.

3. EVALUATION CONDITIONS

Initial evaluation is being performed using the TI-105 word speech corpus. [5] This corpus includes speech spoken in various taking styles, includes 8 speakers, (5 male and 3 female) and provides 5 training tokens and 2 testing tokens per condition for each vocabulary item.

Noise was added to the clean speech condition to evaluate performance under noisy channel conditions. Noise was speech babble recorded in a public meeting place with many background speakers. For this evaluation we used Ghitza's definition of signal to noise ratio [4] as the ratio of the energy per speech sample in the clean speech and the noise averaged over the entire duration of the utterance.

The recognition system was a word-based HMM system with eight speech states per word model, continuous density observations and a single tied diagonal covariance matrix for every state. This robust recognizer provides low error rates on many isolated-word databases.

Both auditory front ends produce high-dimensional feature vector outputs. For classification, the dimensionality was reduced using the same inverse cosine transform used for the mel-scale cepstral front end. For these purposes, all auditory model outputs were treated as representing spectral magnitude. This was done in lieu of a more advanced data reduction technique such as principal components analysis or linear discriminant analysis.

4. RESULTS

Table 1 shows the results of the preliminary recognition experiments. The signal to noise ratio is indicated in the first column, followed by the word accuracy results for the mel-frequency cepstra, (MFC) the "mean rate" response and "synchrony" output (SYN) from Seneff's auditory model, and the results using the EIH model. The binomial standard deviation of the word accuracy rate assuming the MFC performance level is indicated in the final column.

These preliminary results are encouraging. There is no degradation in performance at low noise levels, and all front-ends provide reduced error rates at both intermediate and high noise levels.

SNR	*MFC*	MEAN RATE	SYN	EIH	σ
clean	.5	.7	.8	.8	.2
30	.6	.8	.8	.7	.2
24	1.1	.8	.7	.7	.3
18	2.7	2.1	2.0	2.1	.4
12	8.4	7.6	7.4	7.4	.7
6	26.9	22.9	21.6	22.0	1.0

Table 1: Word Accuracy Rates for Various Signal Representations

5. FUTURE WORK

These results are preliminary, and more experiments are planned to further investigate the effect of these and other signal representations on the performance of speech recognition systems. These include:

1. Evaluation of additional signal representations
2. Exploration of data reduction techniques such as principle components analysis and linear discriminant analysis
3. Exploration of techniques such as the combined use of Seneff's "mean rate" and "synchrony" outputs
4. Evaluation of auditory models on continuous speech using the DARPA Resource Management (RM) corpus

REFERENCES

[1] Davis, S. B. and Mermelstein, P., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, August 1980.

[2] Seneff, S., "A Computational Model for the Peripheral Auditory System: Application to Speech Recognition Research," *Proceedings of ICASSP-86*, April 1986.

[3] Seneff, S., *Pitch and Spectral Analysis of Speech Based on An Auditory Synchrony Model*, PhD Thesis, Massachusetts Institute of Technology, January 1985.

[4] Ghitza, O., "Auditory Nerve Representation as a front end for Speech Recognition in a Noisy Environment," *Computer Speech and Language*, 1986.

[5] Rajesekaran, P. J., Doddington, G. R., and Picone, J. W., "Recognition of Speech Under Stress and in Noise," *Proceedings of ICASSP-86*, April 1986.