APPLICATION OF AN ADAPTIVE AUDITORY MODEL TO SPEECH RECOGNITION

Jordan R. Cohen

Continuous Speech Recognition Group, IBM T. J. Watson Research Center, Yorktown Heights, N. Y., U. S. A. (Current Address: Institute for Defense Analyses, Thanet Road, Princeton, N. J. 08540, U. S. A)

ABSTRACT

An adaptive model of the firing rates found in the auditory nervous system was configured as a signal processor for the IBM speech recognition system. The signal processor was tested on sentences drawn from office correspondence. Several experiments were done in low noise office environments using various microphones and different speakers. The system performance improved substantially compared to performance using a standard signal processor.

INTRODUCTION

Speech recognition systems sample speech signals with a signal-processing front end. One school of thought suggests that an auditory model is the 'ideal' signal processor for such applications, but performance figures available to date do not support the choice of auditory models over more standard signal analyses. This note reports the development and testing of a signal processing algorithm based on some aspects of the mammalian auditory system.

COMMENTS ON THE IBM SPEECH RECOGNITION SYSTEM

Information about the IBM speech recognition system is widely avaliable (Bahl, Jelinek and Mercer, 1983; Nadas, et. al., 1981). The 5000-word vocabulary isolated word dictation system developed at IBM was designed from a communications theory view of speech recognition. It is assumed that a talker formulates a complete English sentence and transforms it into a noisy acoustic signal. This acoustic signal is then captured by an acoustic processor which produces a series of (vector quantized) labels, discrete in both time and identity, from which a decision is made about the most probable sentence given the acoustic input. The probabilistic implementation of the system allows training of the linguistic decoder, but the system performance depends on the reliability of the acoustic processor.

The acoustic processor consists of two sub-systems. A signal processor transforms the high-bandwidth speech signal into a vectorized time signal sampled at a modest rate, and a labeller quantizes the resultant vectors once each centisecond. The standard system uses 30 filter-bank energies once each centisecond as its signal processor, and labels are assigned on a minimum Euclidian distance basis relative to prototypical vectors derived from training data. The signal processor reported here replaces the filter bank with an auditory model.

THE MODEL

The auditory model consists of a frequency analysis followed by perceptually motivated scaling and nonlinear adaptation. The frequency analysis is performed by a 20-band filter bank whose center frequencies and bandwidths correspond closely to those of auditory critical bands (Zwicker, Flottorp, and Stevens, 1957), roughly modeling the selectivity of the auditory system. A compressive power-law transformation is applied to the output from each filter, approximating loudness scaling (Stevens, 1955) and reducing the variability of the vector signal as compared with the original. The compressed signals form the inputs to a reservoir-type model of neural firings (Schroeder and Hall, 1974) which relates stimulus intensity to auditory-nerve firing rate, and which captures certain of the onset and offset characteristics of the neural response.

SIGNAL ACQUISITION AND FILTERING

Speech is captured using a far-field desk-mounted microphone (PZM-6). The speech signal is bandpass filtered (180 Hz to 8 kHz), and is digitized. Power spectra are computed with an FFT. A critical band filter bank is approximated by summing the squared Fourier coefficients (intensity) in each of 20 non-overlapping bands spaced one critical band apart.

The output of each filter is converted from intensity to loudness level by mapping each output power to its equivalent based on the Fletcher-Munson curves (Fletcher and Munson, 1937) and an estimate of the gain of the acoustic system. A conversion to loudness is performed by taking the third (in practice, the fourth) power of the output energy, and scaling such that 40 dB = 1 sone.

SHORT TERM ADAPTATION

Following the lead of Schroeder and Hall (1974), short term adaptation is modeled by assuming the existence of a reservoir holding some amount (n) of neurotransmitter. The change in the amount of neurotransmitter available at time t is described by

$$dn/dt = A - (S_0 + S_H + Dq)n(t).$$

A, D, S_0 and S_H are constants (estimated from psychophysical data), q is the square root of the loudness from each filter, and n is an internal state associated with each filter. This equation states that the change in neurotransmitter is equal to the replacement rate A minus the product of the amount of neurotransmitter available at that time with the sum of the spontaneous rate constant S_0, a decay constant S_H, and a scale D times the square root of the input loudness. The firing rate of that channel is expressed as

$$f = (S_0 + Dq)n(t).$$

These transformations were incorporated into the test system, and the output of the signal processor was substituted for the filter bank outputs of the previous standard process (Das, 1983).

RESULTS

Four talkers recorded the standard 100-sentence training corpus, and then recorded a 50-sentence test corpus at a later time. Signal processing was done twice, once using the filter bank and a second time using the auditory model front end. The system was trained for each speaker using the standard forward-backward algorithm. Results were as follows:

Table 1. Error rate and decoding times for four speakers using two separate front end processes. FB = Filter Bank, AM = Auditory Model.

	Error rate for 50 sentences (%)		*Decoding time (min)*	
Speaker	FB	AM	FB	AM
JRC	6.3	4.7	77	48
FRJ	7.9	4.4	75	38
LRB	4.2	2.3	43	32
PAF	6.6	4.0	99	61
Average	6.3	3.9	74	45

Error rates are expressed as the percentage of incorrect words in the entire test corpus, counting homophones of the correct word as incorrect. Decoding time is the time for the search through the possible sentences, and does not include signal processing time, labelling, clustering, training, and other overhead. Both error rates and decoding times are significantly lower using the auditory model than using the standard filter bank. The overall error rate is reduced by 40 percent. Informal experimentation using different speakers and microphones confirmed the efficacy of the new front end. Several of these experiments are summarized in Table 2.

Table 2. Decoding error rates for various speakers and two microphones. All experiments were trained on 100 sentences of training data, and tested on 20 sentences of test data (299 words). The test text was the same in each experiment. ER = Error Rate (%)

Speaker	Microphone	ER	ER
RLM	lip	3.6	3.3
RHR	lip	7.0	4.6
MAP	lip	6.0	3.3
RLM	lavalier	22.0	2.6
MAG	lavalier	9.3	6.0

The lip microphone was a Sure SMS-10, mounted near the corner of the talker's lips, and the lavalier microphone was a dynamic mike hung from a standard lavalier mount. The word error rates decreased for every speaker, although the decrease for RLM using a lip mike is quite small. (Some of the errors in this corpus are "language model" errors, in that the word strings are highly improbable given our particular 5000 word trigram model. Thus it is extremely difficult to demonstrate error rates below 2 percent for this corpus and language model.) The reduction from 22 percent error to 2 percent error for RLM's recordings using the lavalier microphone is quite striking, but in a different series of experiments using only long- term adaptation, the error rate on this corpus was decreased to 5 percent; much of the decrease is due to gain normalization. Decoding times were always less using the new front end than with the previous signal procesor.

Speakers MAG and PAF are both female, the rest of the speakers in the experiments reported here are male. No consistent difference has been noted in our recognition results between male and female speakers.

SUMMARY

A simple auditory model was developed and tested as a signal processing system for the IBM speech recognizer. It decreases the number of errors made by the system by approximately 40 percent in controlled tests.

ACKNOWLEDGEMENTS

I am indebted to Raimo Bakis, with whom this concept was originally developed, and to Michael Picheny, Robert Mercer, and Fred Jelinek for their support. Lou Braida aided me in debugging my explanations, and David Nahamoo's cooperative competition was of great aid.

REFERENCES

1. Bahl, L. R., Jelinek, F., and Mercer, R. L. A maximum likelihod approach to continuous speech recognition. *IEEE Trans. on pattern analysis and machine intelligence.* PAMI-5, 1983, 179-190.

2. Das, S. K. Some dimensionality reduction studies in continuous speech recognition. *IEEE Conf. on Acoust., Speech, and Signal, Proc.* 1983, 292-295.

3. Fletcher, H. F. and Munson, W. A. Loudness, its definition, measurement and calculation. *J. acoust. Soc. Am.* 5, 1937, 82-108.

4. Nadas, A., Mercer, R. L., Bahl, L. R., Bakis, R., Cohen P. S., Cole, A. G., Jelinek, F., and Lewis, B. L. Continuous speech recognition with automatically selected acoustic prototypes obtained by either bootstrapping or clustering. *IEEE Int. Conf. on Acoustics, Speech, and Sign. Proc.* 1981, 1153-1155.

5. Scharf, B. (1978). "Loudness," in Carterette and Friedman (eds.) Handbook of Perception, Vol. IV, Academic Press, p. 180-242.

6. Schroeder, M. R. and Hall, J. L. Model for mechanical to neural transduction in the auditory receptor. *J. acoust. Soc. Am.* 44, 1974, 1055-1060.

7. Stevens, S. S. The measurement of loudness *J. acoust. Soc. Am.* 27, 1955, 815-829.

8. Zwicker, E., Flottorp, G., and Stevens, S. S. Critical band width in loudness summation *J. acoust. Soc. Am.* 29, 1957, 548-557.