AUDITORY MODEL INTERFACES INTO A DTW RECOGNISER

Daniel Woo*, Phillip Dermody*, Richard Lyon$^{\pi}$ and Bruce Lowerre†

*Speech Communication Research
National Acoustic Laboratories

$^{\pi}$Advanced Technology Group
Apple Computer, Cupertino CA.

†Talk Is Cheap, Inc.

ABSTRACT - Auditory models have been proposed as one way to improve the robustness of current speech recognisers. In this work the Lyon auditory model is coupled to a DTW speech recogniser and comparisons are made between LPC coefficients and autocorrelation coefficients derived from the auditory spectrum. The results for both are compared using speaker dependent recognition for three speakers across different signal to noise ratios. The results suggest that auditory models can be used for interface to current recognisers and that the autocorrelation output plus a Euclidean distance measure provides the best performance in the current configuration.

INTRODUCTION

The purpose of this study is to assess the performance of an auditory model as a front-end processor in a speech recognition system. Auditory models are emerging as a novel preprocessor for speech recognition systems but the marriage of the auditory model to current speech recognisers is still in its infancy. This paper describes the performance of an auditory model-based recognition system using two different techniques and compares the results with a standard linear prediction (LP) technique.

Conversion of the speech signal into a parametrized form is illustrated in Figure 1. The speech signal is processed by the auditory model, producing a spectrum-like slice at 5 msec intervals. The output of the model has many channels and it is necessary to reduce the spectrum to a smaller dimensional vector which retains the salient information of the utterance. This vector is used in the pattern recognition process to identify the closest matching stored template.

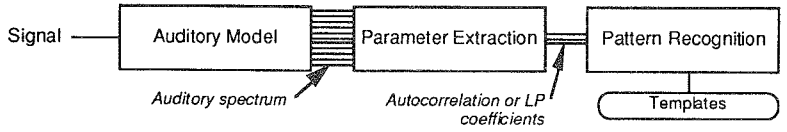

Figure 1. Auditory model interface into a pattern recogniser.

Auditory Models

Auditory models are physiologically motivated algorithms which model the human auditory system and have been reported to improve the recognition performance in noisy environments (Ghitza, 1987).

The auditory model used in this study (Lyon, 1982. Slaney, 1988) produces 84 channels which represent the average neural firing rate at a particular place along the basilar membrane. The cochlear model simulates the filtering characteristics of the ear and non-linear operation of the hair cells. Four distinct operations can be identified in the Lyon model (Figure 2). The signal is attenuated by 20dB

prior to entering a cascade of second order filters. Two of these filters account for the spectral shaping imposed by the outer and middle ear while the remaining filters model the frequency sensitivity of the basilar membrane at different positions along its length. Filters which have centre frequencies above 1 kHz have been designed with a Q of 8. The remaining filters, with centre frequencies below 1 kHz have bandwidth approximating 125 Hz.

Conversion of basilar membrane motion into neural activity by the inner hair cells is modelled by half wave rectifying the output of each filter. Four stages of cross-coupled automatic gain control are used to model adaptation and inhibition effects. The model can produce an 84-channel vector for each sample point. To reduce the information rate, the auditory model output is decimated, producing a spectrum at intervals corresponding to 80 input sample points. A smoothing operation is used to reduce aliasing after decimating and is achieved by passing each channel of the auditory model through two low pass filters with cutoff frequencies at 10.6 Hz.

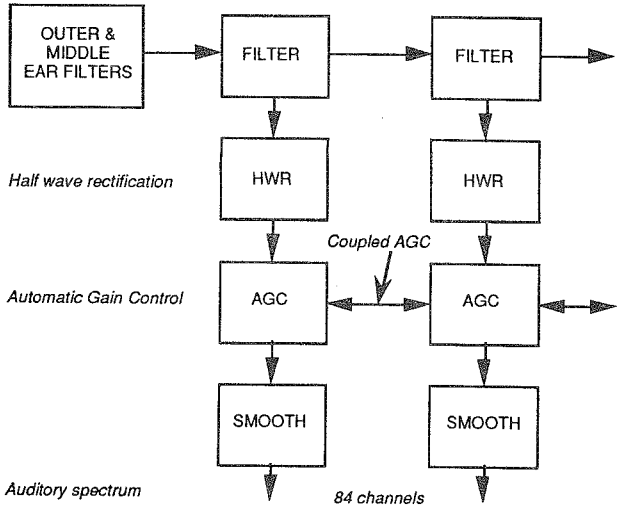

Figure 2. The Lyon Auditory Model.

The auditory output for the digit 'seven' is shown in Figure 3 for three signal to noise ratios. The output of the first auditory filter which corresponds to a position near the base of the basilar membrane is located at the top of the vertical axis. Prominent features of the utterance are maintained in noise even though the dynamic range is substantially reduced.

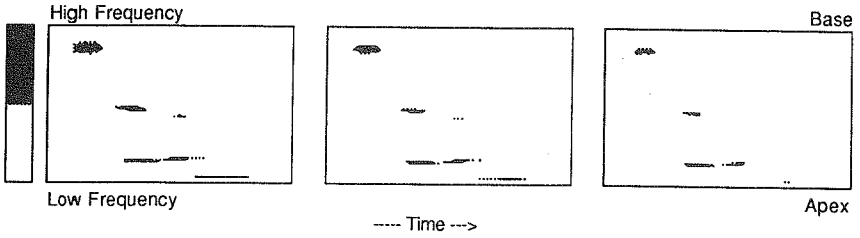

Figure 3. Auditory spectrogram for the digit 'seven'. (a) No noise, (b) 20dB and (c) 10dB.

The auditory model information bandwidth is too great to couple directly into our DTW speech recogniser. A reduction in the bandwidth is achieved by reducing the dimensionality of the output vector while attempting to retain the salient speech features.

Parameter Extraction

Two strategies for reducing the speech signal to a lower dimension feature vector considered in this study are the derivation of LP parameters and autocorrelation coefficients from the auditory spectrum.

If it is assumed that the squared output of the auditory spectrum is a power spectrum then an inverse Fourier transformation can be used to produce a set of autocorrelation coefficients. Ten LP coefficients are calculated by applying Durbin's recursive algorithm (Makhoul, 1975) on a normalised autocorrelation vector.

Figure 4 compares spectra for the vowel /or/ derived from an autocorrelation vector and LP coefficients using only 10 coefficients.

The Itakura distance is used to measure the similarity between LP vectors and a Euclidean distance is used for the auditory autocorrelation coefficients.

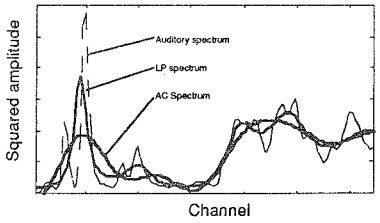

Figure 4. Comparison of an auditory spectrum with LP and autocorrelation spectrum.

Pattern Classification

A dynamic time-warping (DTW) pattern classifier returns the minimum distance between two templates when the time relationship between the reference and test template is warped under constrained conditions (Figure 5). We have used a DTW recogniser to evaluate the performance of the auditory model.

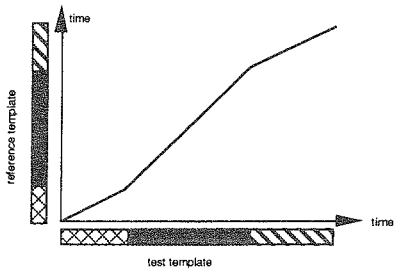

Figure 5. Warping of the time relationship between test and reference templates.

Each template contains a chronological sequence of vectors (found during the parameter extraction phase) which represent the acoustic events of the speech signal. If two feature vectors representing different instances of the same word are compared it will be observed that the time alignment between

similar features differs even though the sequence in which events occur might be the same. If it is permissible to stretch and compress the duration of these events by a restricted amount a better measure of similarity is possible.

Database

An isolated digit recognition task was used to compare the different configurations. In order to train the system ten training utterances were required to generate the reference templates. An eleven digit database which consisted of ten repetitions of each of digit (oh, one,..., nine, zero) was recorded by three male speakers in quiet conditions at 16 kHz sampling frequency. Noise-corrupted samples were produced by mixing one segment of pink noise with all the digits at 20 and 10dB SNR (rms). Training and testing were performed on the same database.

RESULTS

The speech recognition system was trained with noise-free speech and tested using utterances in no noise, 20dB and 10dB SNR (rms). The autocorrelation coefficient and Euclidean distance measure performed better in noisy conditions compared with the auditory LP method and Itakura distance. All methods produced similar recognition scores in clean conditions. In comparison with a standard LP system, recognition performance gains could only be observed at 10dB.

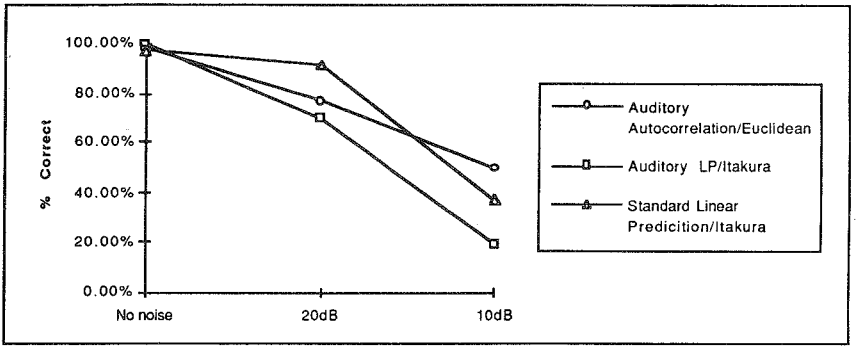

Figure 6. Recognition performance.

It is unclear why the performance of the auditory model autocorrelation method is poorer than standard LP at 20dB and better at 10dB. We are currently investigating whether this phenomenon is the result of the parameter selection process or from a conflict between the auditory feature vectors and template creation process in the DTW.

CONCLUSIONS

In the current configuration we have shown that autocorrelation coefficients extracted from the auditory spectrum combined with a Euclidean distance measure can be used as a simple interface into a DTW recogniser. Additional processing of the autocorrelation vector to form linear prediction coefficients combined with an Itakura distance measure does not improve recognition performance in noise.

We are currently reviewing other methods of data reduction suitable for interfacing auditory models to current pattern classifiers which enhance speech recognition performance.

REFERENCES

Ghitza, O. (1987) *Robustness against noise: The role of timing-synchrony measurement.* Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Dallas Texas.

Lyon, R.F. (1982) *A Computational Model of Filtering, Detection, and Compression in the Cochlea,* Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, Paris, May 1982.

Makhoul, J. (1975) *Linear Prediction: A Tutorial Review,* Proc. IEEE, Vol. 63, No 4, 561-580.

Slaney, M. (1988) *Lyon's Cochlear Model.* Apple Computer Technical Report #25, (Apple Computer Inc.).

DETECTION OF WORD BOUNDARIES IN CONTINUOUS HINDI SPEECH USING PITCH AND DURATION

G.V.RAMANA RAO
Department of Computer Science and Engineering
Indian Institute of Technology
Madras 600 036, INDIA.
Email : ramana@iitm.ernet.in

Reliable detection of word boundaries in continuous speech is an important problem in speech recognition. Many studies established the importance of prosodic knowledge in detecting word boundaries. In this paper we report a word boundary hypothesisation technique based on the durational knowledge for Hindi. Recently another technique using pitch patterns was proposed for Hindi. We have also shown in this paper that combining the duration and pitch knowledge leads to significant improvements in the overall detection of word boundaries.

1. INTRODUCTION

Word boundary hypothesisation is an important problem in continuous speech recognition. If word boundaries can be recognised accurately, many of the techniques developed for Isolated Word Recognition(IWR) systems can be adapted for continuous speech recognition. In most of the speech recognition systems built, the word boundary hypothesisation is done as part of the lexical analysis. However, in absence of word boundaries, the dictionary search requires large amounts of computer storage and time. Studies on English(Harrington and Johnstone, 1987) showed that in absence of word boundaries the number of possible word strings matching an utterance at a mid class level, may exceed 10 million. Similar results were obtained for Hindi by us. Hence in continuous speech recognition systems it becomes necessary to perform word boundary hypothesisation before performing lexical analysis.

Despite its importance, the problem of word boundary hypothesisation has only recently attracted attention. Harrington (Harrington, Watson and Cooper, 1989) examined the use of phoneme sequence constraints for word boundary detection in English, but in the presence of ambiguities in the phonemes(as in the context of speech recognition), he found them to be of limited use. However, he found that a strong/weak classification of syllables can lead to a good word boundary detection. For Hindi, Ramana Rao (Ramana Rao and Yegnanarayana, 1991) reported the use of some language clues for hypothesising word boundaries. Based on simulation studies, he reported that these clues can be used at fairly large error levels in the phoneme strings. In a recent paper, Ramana Rao (Ramana Rao, 1992) also reported the use of some spectral clues to detect word boundaries. Rajendran and others (Rajendran, Madhu Kumar and Yegnanarayana, 1992) also reported a word boundary detection technique using pitch patterns for Hindi. In this paper we report our results on using duration knowledge to hypothesise word boundaries.