PERFORMANCE OF A PERIPHERAL AUDITORY MODEL ON PHONEMES IN COMBINATION

A. Samouelian and J. Vonwiller

School of Electrical Engineering,
The University of Sydney

ABSTRACT - A peripheral auditory model, implemented as a front-end speech processor, has the potential to provide relevant acoustic features for speech recognition. The model's performance on utterances containing 'l' and 'r' sounds in various positions in connected speech is described. The effects of stress on these consonants is also examined. Results on real speech signals are presented.

INTRODUCTION

A computational model of the peripheral auditory model based on the processes of the cochlea has been under development at Sydney University for the past year [Samouelian, 1989]. The model was developed for two primary reasons. Firstly, the model exhibits a high degree of structured regularity and process concurrency, thus lending itself to efficient ASIC implementation. Secondly, the model will be used as a front-end signal processor for an acoustic feature extraction unit currently under development at the university. The ultimate aim is to develop a robust speech recogniser, which can reliably produce phonetic labels and operate successfully on continuous speech signals.

An essential part of any speech recognition development system is the data base collection. Work has commenced to collect such a data base. We have currently over 1000 consonants in syllable initial and syllable final positions, and over 150 consonant clusters, spoken by an Australian male, female and child speakers, and recorded onto a Digital Audio Recorder (DAT). These consonant combinations are currently being processed by the auditory model. The data base will help in the development of a set of acoustic feature extraction algorithms based on the features generated by the model.

In rule based speech recognition, the reliable identification and fine class classification of the liquids 'l' and 'r' have been particularly difficult to achieve, because of the high degree of spectral and temporal variability of these phonemes. Recognition is made more difficult by the wide allophonic variability exhibited by these consonants. This view has been reinforced by other researchers working in this field.

As an alternative approach to attempting to resolve this problem, an experiment was conducted to examine the benefits of applying the auditory model to enhance the spectral and temporal characteristics assosiated with these consonants.

Detailed descriptions of the auditary model have been published elsewhere (Samouelian & Summerfield, 1988, 1989) and (Samouelian, 1990). Here, only a brief description is presented. The model is shown in figure 1. It consists of 64 overlapping critical band filters, spanning a frequency range of 100 to 6310 Hz and spaced linearly on the Bark scale at 0.3 Bark. The bandwidth of each filter is set at 0.5 Bark. The linear filterbank is followed by the non-linearities, consisting of compressive rectifiers, adaptors and automatic gain controls, which simulate the transduction stage of the cochlea. The final stage is fed through a synchrony detector, which enhances vowel recognition.

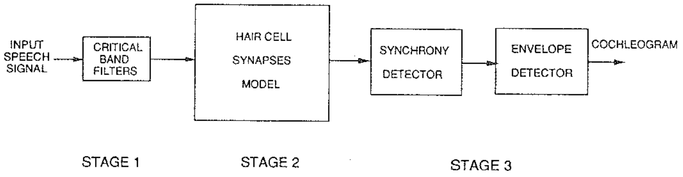

Figure 1: The peripheral auditory model

The model was simulated using a set of generic signal processing modules, written in C. All the modules share a common generic communication protocol, which enables a structured model of the cochlea to be constructed, using the UNIX *piping*, *redirection* and *tee* facilities. The auditory model has been installed within AUDLAB, an Interactive Speech Signal processing Software Package, which is installed on a Masscomp 5520 computer and ESPS, Entropic Signal Processing System, which is installed on a Sun 4 computer.

DATA SELECTION

Vowels can be described sufficiently by their formant structure but the consonants have to be characterized by more complex features which reflect temporal as well as spectral properties (Suen & De Mori, 1982). Liquids are similar to vowels but use the tongue as an obstruction in the oral tract, causing air to deflect around the tip or dorsum. The liquid 'l' is a voiced lateral alveolar consonant. The tip contacts the alveolar ridge and divides the airflow into two streams on either side of the tongue. Voiced lateral alveolar consonants are characterized by the existence of voicing and a domination of low frequency energy in the F1 and F2 formants. The positioning of the articulators in 'l' causes very high F3 which is outside the range typical of vowels. Partitioning of the air stream causes a spectral zero to apper near 2 kHz. Temporal discontinuiuties caused by tongue tip breaking contact with the alveolar ridge may also be useful for fine class classification.

The liquid 'r' is a voiced retroflex alveolar consonant. As with 'l', this is also characterized by domination of low frequency voiced energy. English 'r' causes F3 to descend much lower than for any other phoneme. The formant transitions for both these consonants are smoother and slower than for other consonants. In clusters a stop consonant 'p' will have increased aspiration if followed by the 'r' consonant (O'Shaughnessy, 1987).

The sound combinations were selected to contain the liquids in several combinations in connected speech. Four utterances (two sentences and two phrases) were recorded with various stress patterns, each stressing a different word in the utterance. The utterances were spoken carefully in Educated Australian English.

A female speaker was recorded saying the following sentences:
(a) "Bill's splint is off".
(b) "Roy will really pray aloud".
A male speaker was recorded saying the following phrases:
(c) "Grey plane, grape lane".
(d) "She prayed, sheep raid".
The (b) sentence was repeated several times, each time with the stress on a different word.

The sentences allow for the analysis of 'r' in the initial, singleton and cluster positions and the 'l' in the initial, medial, final, singleton and cluster positions. Using these samples we investigated the two consonants for:
(a) basic segmental features.
(b) segmental features related to position in word.
(c) segmental features related to syllable boundaries.
(d) segmental features related to stress.

PERFORMANCE EVALUATION

The performance of the model is shown in figure 2 for the phrase "Grey plane, grape lane", and figure 3 for the phrase "She prayed, sheep raid". Figures 2(a) and 3(a) show the respective traditional wideband speech spectrograms calculated using 256 point FFT. Figures 2(b) and 3(b) show the respective cochleograms.

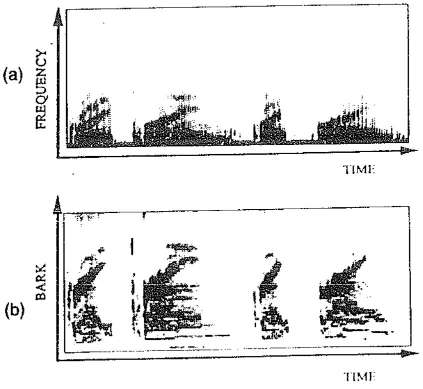

Figure 2: Performance of the model on the phrase "Grey plane, grape lane"

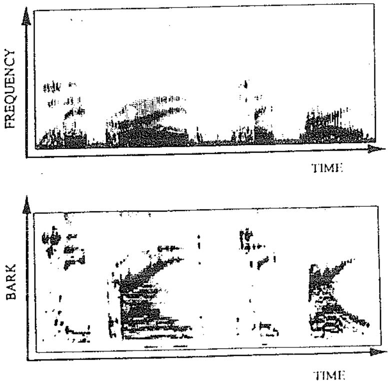

Figure 3: Performance of the model on the phrase "She prayed, sheep raid"

Figures 4, and 5 show the performance of the model on the sentences "Bill's splint is off" and "Roy will really pray aloud" respectively. Figures 4(a) and 5(a) show the wideband spectrogram, and figures 4(b) and 5(b) show the respective cochleograms. Figures 6(a), 6(b), 6(c), 6(d) and 6(e) show the output of the model for the stress patterns on the sentence "Roy will really pray aloud". The words were stressed consecutively.

Basic segmental features.

From the figures, both 'l and 'r' produce the formant structure described above, but the clarity of the picture was influenced by the position in the word.

Word position.

Initial prevocalic 'r' in 'Roy' in figure 5 was signalled by distinctive steeply rising F3, inferring an initial low F3 during the consonant followed by a rapid rise during the transition into the following vowel. The output from the cochlea model show this particularly well.

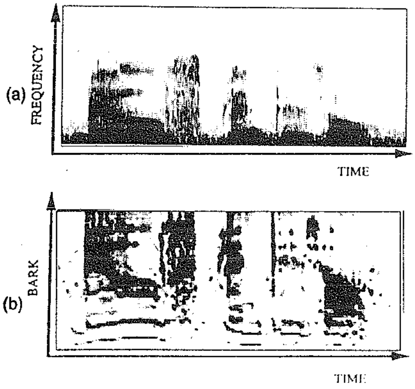

Figure 4: Performance of the model on the sentence "Bill's splint is off"

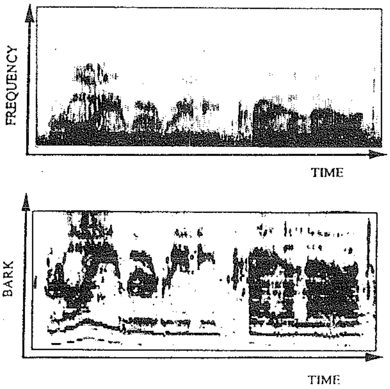

Figure 5: Performance of the model on the sentence "Roy will really pray aloud"

In figure 2, initial 'l' in 'lane' demonstrates the discontinuities expected and a weak high F3. Medial 'l' in 'really' in figure 5 was signalled by rising F3 and a falling F2 and a discontinuity associated with the amplitude of these formants. This again distinguishes the 'l' and 'r'. Final 'l' in 'will' in figure 5 had fewer discontinuities than elsewhere as there was no articulatory break. The final 'l' in 'will' is followed by 'r' in 'really' in figure 5 which pulls the F3 down steeply to initiate that sound in all but one of the samples in this sentence. The one sentence where it does not occur is in figure 6(b), where the final 'l' is in a stressed syllable. Final cluster 'l' in 'Bill's' in figure 4 also has no discontinuities in transitioning from the vowel sound though it breaks completely before the combined 's' of 'Bill's

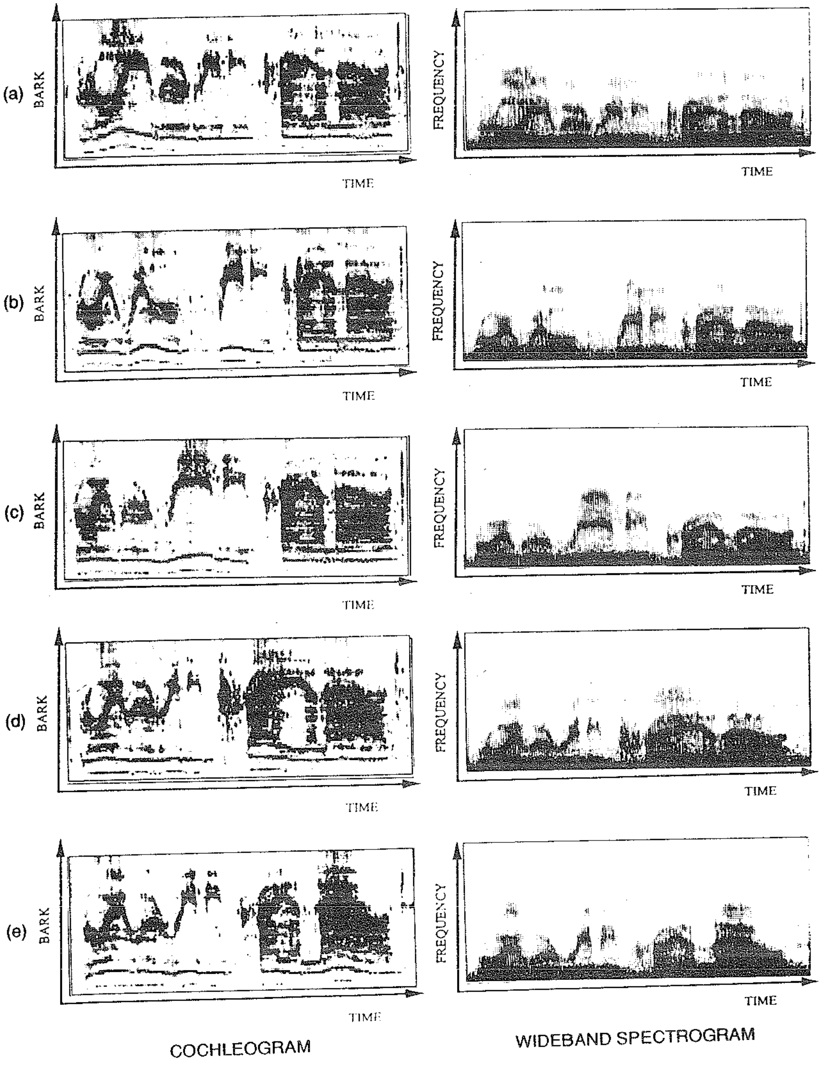

Figure 6: Performance of the model on the sentence
"Roy will really pray aloud"

splint'. The F3 is not affected by the following 's'.

Syllable boundaries.

Figure 2 shows that with 'l' an initial singleton in 'lane', the central portion has a longer duration than the 'l' in an initial cluster ('plane'). F3 is weak but high and there are some discontinuities in the 'l' singleton which are less evident in the 'l' cluster. In the cluster in 'prayed', the 'r' in figure 3 became fricated following the 'p', and F3 was lowered less by the following 'r'. This distinguished the 'prayed' and 'raid' syllable onsets.

Stress effects.

In figure 2, initial 'l' in 'lane' has a longer duration and a more obvious F3. In the final position 'l' in 'will' in a stressed word retained it distinguishing features, whereas in unstressed syllables it had not. Stress for 'r' consonants enhanced the F3 transitions in figure 6(a).

CONCLUSION

This paper evaluated the performance of a peripheral auditory model on classification of consonants in continuous speech. The aim of this work was to examine the application of the auditory model being developed at Sydney University to enhance recognition of the liquids, 'l' and 'r'. The model was able to identify both consonants in the syllable initial position. In the final position the 'l' was most readily identifiable when in a stressed syllable. In the phrases, syllable boundaries of the kind 'C CV' and 'CCV' were recognised. The cochlea model enhances the transitions from one sound to another, making transitions for 'l' and 'r' more distinct. We also examined the effects of stress on these consonants. Our ultimate aim is to establish a catalogue of auditory model output for phonemes in combination to help us determine a set of features that need to be extracted to represent the particular phonemes. A further motivation behind this work is to take some steps towards establishing an Australian data base to assist in developing an acoustic feature extractor, using the auditory model as a preprocessor.

ACKNOWLEDGEMENT

We thank Dr. Clive Summerfield for his constructive suggestions during the writing of this paper.

REFERENCES

O'Shaughnessy D. (1987). "Speech Communication, Human and Machine", Addison-Wesley publishing company.

Samouelian A. & Summerfield C. D. (1989). "Front-end speech signal processor for speech recognition", Proc. IREECON89 Int. Conf., September 1989, Melbourne, Australia, pp 112-115.

Samouelian A. & Summerfield C. D. (1988). "Computational model of the peripheral auditory system for speech recognition: Initial results", Proc. Second Australian Int. Conf. on Speech, Science and Technology, November 1988, Sydney Australia, pp 234-239.

Samouelian A. (1990). "Speech recognition front-end using auditory model", to be published in Int. Conf. on Signal Proc. '90 proceedings, 22-26 Oct., 1990, Beijing, China.

Suen C. Y. & De Mori R. (Editors) (1982). "Computer analysis and perception", CRC press Inc. Volume 2, page 24.