Proceedings of The Institute of Acoustics 4A21

THE INPUT STAGE TO A MODEL OF THE SPEECH PERCEPTIVE SYSTEM

M. WEST, N. HADDOCK, S. G. STRADLING, J. B. COPAS

SPEECH RESEARCH GROUP, UNIVERSITY OF SALFORD

Introduction

A model of the speech perceptive system can be considered simply as a speech recogniser with two important constraints which are not imposed on other recognisers. Firstly, it may only incorporate "devices" which are known to be used in the human system and, secondly, it must use those devices in the same manner as the human system.

A first stage is being constructed to some extent in ignorance of the detailed form of subsequent stages since information about the higher levels of speech processing is scanty. It is, however, our intention to move on to develop additional "stages" within the above two constraints. A degree of generality is therefore required in terms of control inputs to the devices in order to permit "top down" controls which will be necessary in the ultimate heterarchical system.

Design Objectives

The objective of this first stage is to produce a set of "black-boxes", each of which has a function for which there is some experimental evidence in the human system. It is particularly important to include all possible "black-boxes" and not to combine more than one function within a given box, even if there is some evidence for this on occasions during perception. In other words, the boxes may not have a one to one correspondence with the neural pathways, but they do have a one to one correspondence with the functions of these pathways.

This first stage, when used in recognisers, is often referred to as acousto-phonemic (1). However, our first stage cannot make such an ambitious claim and would be better described as acousto-acoustemic (2), in that the acoustic signal is converted to a prephonemic structure.

In order that the model should be accurate, the detailed functions of the acousto-electrical transduction within the ear might be thought necessary. However, the black box functions themselves are designed to incorporate these features in a more relevant form to the subsequent processing of the input speech. For example, the precise nature of the frequency analysis that occurs in the Basilar Membrane and its subsequent conversion into different neural responses (3) is not vital to this first stage and is effectively incorporated within the systems discussed.

The Processors

For each of the processors the references give the experimental justification. In all cases this justification results from psychophysical investigation. The implementation of the boxes is also considered in terms of digital computer software. Both the justifications and the software have a considerable literature and only key references are mentioned.

THE INPUT STAGE TO A MODEL OF THE SPEECH PERCEPTIVE SYSTEM

Pitch Extractor

We are only concerned here with the extraction of the speech fundamental frequency, Fx, and not the general ability of the ear to extract the pitch of a complex signal. In this case, it is the facility of obtaining subjective pitch from a number of harmonics of Fx rather than being able to perceive Fx itself that is of interest. Very often, Fx is perceived even when it is not physically present. Plomp and Smoorenburg (4) give extensive evidence for this function.

Implementation of a pitch extractor with speed and reliability is by no means easy (5). A number of pitch extraction techniques are being examined at Salford, including one based on a technique developed by Moorer (6) which uses a variable comb filter.

Fine Time Analyser

Zwislocki et Al (7) have shown that neural bursts depend on the slope of the input acoustic signal. Green (8) has shown that the ordering of signals can be discriminated down to as little as 1-2ms. It is also known that vowels may be accurately identified at lengths as short as 27ms (9). Clearly, there are facilities for fine time analysis which are used for speech perception.

The fine time analyser under development produces the following outputs: (i) voiced/non-voiced decision, (ii) slope, (iii) intensity, (iv) VOT (1) is clearly linked to the pitch detector in the later stages of the model.

Frequency Analyser

Pols et Al (10) showed that vowel perception is formant based without any preconceived ideas about formants. They also showed that the first two formants are crucial for vowel identification. It is also known (11) that F3 may be important in cue adaptation in a consonant-vowel situation.

The human perceptor appears to use both bandwidth (12) and relative amplitude information (13) as well as the centre frequency of a given formant. In addition, the transitional changes especially preceding a vowel are of immense perceptive importance (14).

Our frequency analyser must be capable of extracting formant frequency bandwidth and amplitude of the first three formants. There are a number of techniques available for formant extraction which, like those for pitch extraction, have shortcomings. Most of the difficulties arise when the formants are overlapped, or when the vowel is very short (15). Our formant extractor is based on an initial FFT taken over a 10-30ms window. It employs a two stage spectral smoothing technique to extract non overlapped formants. There is no attempt to separate overlapped formants which are viewed as a single merged "formant".

A coarse frequency analysis is necessary to isolate fricatives (16). This is easily implemented using the same system as that used for formant extraction.

The above three systems with their component black boxes give all the required information for subsequent processing in our model. No assumptions as to the nature of that processing or its outcome have been made in their selection.

References

(1) GREEN, P. D. and AINSWORTH, W. A. 1973 Brit. Acoust. Soc. 1973 Spring Meeting, Development of a system for the automatic recognition of spoken Basic English.

(2) KEIL, G. Private communication.

(3) DALLOS, P. 1973 The Auditory Periphery. Academic Press.

(4) PLOMP, R. and SMOORENBURG, G. F. 1970 Frequency Analysis and Periodicity Detection in Hearing. A. W. Sijthoff, 250 - 473.

(5) RABINER, L. R. et Al. 1976 IEEE Trans. ASSP 24, 5, 399 - 418. A Comparative Performance Study of Several Pitch Detection Algorithms.

(6) MOORER, J. A. 1974 IEEE Trans. ASSP 22, 330 - 338. The Optimum Comb Method of Pitch Period Analysis of Continuous Digitised Speech.

(7) ZWISLOCKI, J. J. and SOKOLICH, W. G. 1973 Science, 182, 64. Velocity and Displacement Responses in Auditory Nerve Fibers.

(8) GREEN, D. M. 1973 J. Acoust. Soc. Amer., 54, 373 - 379. Temporal Acuity as a Function of Frequency.

(9) THOMAS, I. B., et Al. 1970 J. Acoust. Soc. Amer., 48, 1010 - 1013. Temporal Order in the Perception of Vowels.

(10) POLS, L. C. W. et Al. 1969 J. Acoust. Soc. Amer., 46, 458 - 467. Perceptual and Physical Space of Vowel Sounds.

(11) BAILEY, P. J. 1974 Perceptual Adaptation of Acoustical Features in Speech. Speech Communication Seminar, 47, Almqvist and Wiksell.

(12) CARLSON, R. et Al. 1974 Acustica, 31, 360 - 362. Two Formant Models, Pitch and Vowel Perception.

(13) AINSWORTH, W. A. and MILLAR, J. B. 1972 Language and Speech, 15, 328. The Effect of Relative Formant Amplitude on the Perceived Identity of Synthetic Vowels.

(14) LIBERMAN, A. et Al. 1967 Psychol. Rev., 74, 431 - 461 Perception of the Speech Code.

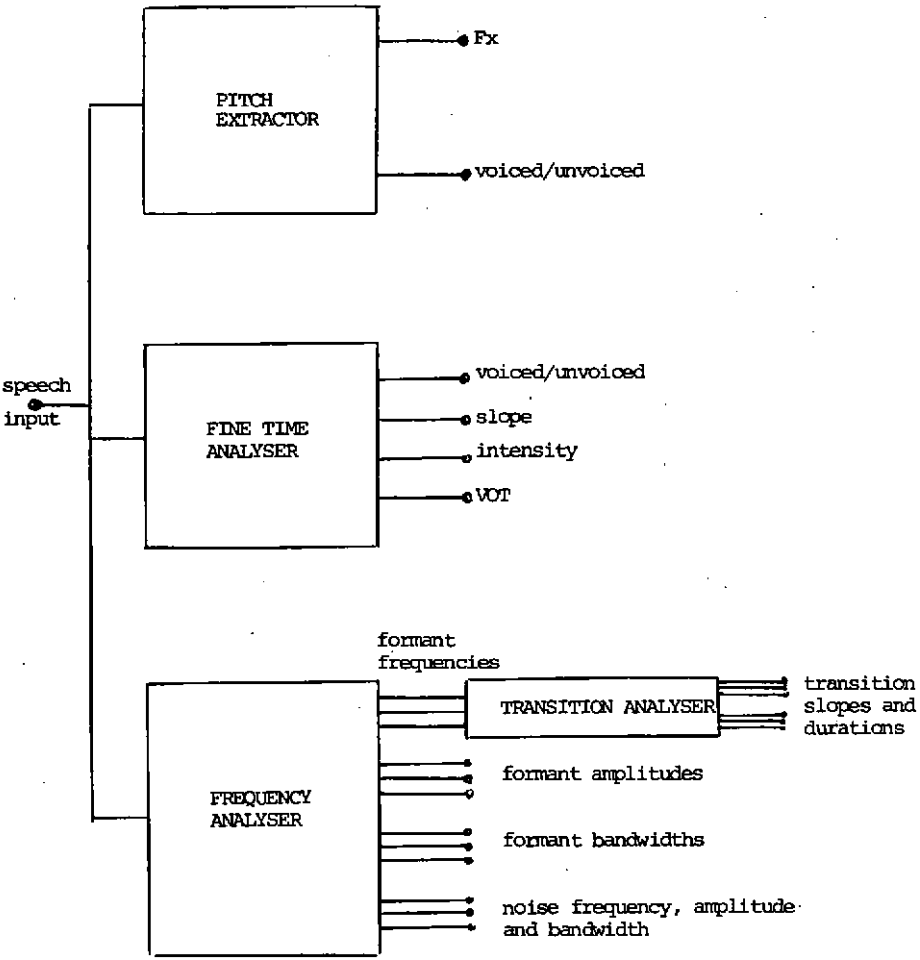
(15) OLIVE, J. P. 1971 J. Acoust. Soc. Amer., 50, 661 - 670. Automatic Formant Tracking by Newton-Rapson Technique.

(16) STREVENS, P. 1960 Language and Speech, 3, 32. Spectra of Fricative Noise in Human Speech.

Acknowledgement

The authors wish to thank Professor F. Knowles (University of Aston) and Mr. R. Johnson (UMIST) for their many helpful suggestions.

THE INPUT STAGE TO A MODEL OF THE SPEECH PERCEPTIVE SYSTEM



Block Diagram of the Parallel Processors for the Input Stage of the Model Speech Perceptor