Proceedings of The Institute of Acoustics

TOWARDS AN AUDITORY PRIMAL SKETCH

Ray Meddis

Department of Human Sciences, University of Technology,
Loughborough, LE11 3TU.

INTRODUCTION

The application of psychological and physiological theories to the development of automatic speech recognisers would appear to be an obvious approach. While this is often suggested, it is rarely implemented, even though this method offers considerable benefits both to the artificial intelligence community and to the development of psychological theories. For example, there has been considerable recent activity in the development of models of mechanical-neural transduction at the hair cell with substantial implications for speech recognition technology. These models are considered more fully elsewhere in these proceedings [1]. The use of band-pass filters equally spaced along a Bark scale is, however, slowly being adopted by the speech recognition community and this brings us one step closer to a physiological approach to hearing.

This paper will look at two ideas drawn from psychological theory. The first involves segmentation of the utterance, while the second concerns the development of an internal model of speech input. The segmentation procedure will draw upon physiological studies of vertebrate vision and work by Marr and co-workers [2] on computational models of the segmentation of visual images prior to the identification of objects. The internal modelling procedure will draw upon Piaget's [3] ideas of how children develop concepts using the principles of 'assimilation' and 'accommodation'.

The recognition device to be demonstrated uses the segmentation procedure to subdivide the auditory input into a small number of segments which are each described in simple terms, to produce a short list of descriptors representing the input. This descriptor list is then used by the modelling procedure to help recognise isolated words and to modify the internal model appropriately in the light of the experience. The recogniser was built primarily to demonstrate clearly the application of these key ideas and has not been optimised for performance. However, there is considerable potential for development in the system which will be explored in future work. Because the data reduction is considerable, the system is suitable for use on micro computers with restricted storage. The system will be demonstrated on a BBC model B computer.

SEGMENTATION

Recent studies in vertebrate visual analysis have shown brain cells which respond exclusively to changes in luminance levels. Marr and co-workers have shown how such units can be used to detect edges in visual stimuli. This edge detection process is used to help compile a "primal sketch" or two-dimensional plan of edges of objects thought to be present in the image. The primal sketch is then the subject of higher order analysis which need not concern us here.

Their work is in two dimensions as befits visual analysis but a simplified one-dimensional analogue will be demonstrated here as suitable for hearing studies. They use luminance levels but the principle is independent of the variable to be analysed, which could be any aspect of the stimulus. Figure 1A shows the progress of a simple zero-cross count for an unfiltered acoustic stimulus being the word "twenty". The discussion below could equally deal with the unfiltered stimulus amplitude, band-pass filtered amplitude or band-pass filtered zero-cross count. It might even accept as input the density of neural spikes generated by a computational model of auditory-neural transduction, [e.g. 4].

The raw zero-cross count shows clear features which can be visually detected and these features such as peaks, flat portions, troughs, etc., are potentially useful units of analysis. The aim of the process to be described is to automatically segment the utterance accordingly. Marr and Hildreth's method smooths the function as a first step. This is shown in Figure 1B. The smoothing is based on moving rectangular windows of width 50 milliseconds. Step two differentiates the smoothed function (Fig. 1C). Note that peaks and troughs in the differentiated function can be used to identify reasonable points of segmentation in the original utterance. Step three, therefore, takes the second order differential (Fig 1D) and notes the points at which the new function crosses the baseline in either direction. By erecting vertical lines from these points, we can see that they form useful segmentation points for the original utterance.

The computer program uses this simple approach although Marr and Hildreth have shown that the use of Gaussian smoothing techniques, when combined with a Laplacian, produce (in the two dimensional situation) optimal results from various points of view. Marr also recommends that various degrees of smoothing be applied and only those segment boundaries used which are indicated by all levels of smoothing. In the current implementation, this complication is ignored while acknowledging that different results occur if the smoothing factor is changed. Figure 2 A-D illustrates the use of this technique for four other words.

DESCRIBING SEGMENTS

The recogniser accepts the segmentation and then seeks descriptors for each segment. The recogniser takes the mean zero cross value between two segment lines as the key descriptor. As a consequence, the word image is stored as a string of values, one byte per segment. The list of descriptors (SEGMENT LABELS) is given to the right of Figs 1 and 2. For most words, the number of arguments is less than five so the data reduction is severe.

Another version of the program supplements the mean zero cross descriptor with the slope in the zero cross function at the initial segment boundary and the length of the segment. These supplementary descriptors are, however, arbitrary and beyond the scope of this paper.

TEMPLATE DEVELOPMENT

A single word recogniser could be devised simply by storing each descriptor string along with the name of the word which gave rise to it. Recognition would then be based on a simple best-match sort through memory. Because only a few bytes are needed per word, this would have only limited implications for storage. Moreover, the search would be restricted to descriptor strings of the same length as the input string, so the search could be only partial. However, the recogniser to be described uses a psychological principal which uses information from each auditory experience to improve the recognition performance but keeps the number of templates to a minimum.

Piaget suggested that concepts, or schemata, grow and differentiate according to biological principles. For example, the concept of a "party" is enriched by every party one attends or learns about. This process of enrichment is called "assimilation". However, only experiences which loosely match the concept can be assimilated to it. Thus, a cycle ride does not enrich the "party" concept. Furthermore, some experiences could be classified as "parties" but because of significant differences eventually call for a new concept. Thus, a conference may be similar to a party from many points of view but the differences require that an additional concept for conferences is formed. This process of differentiation of the conceptual structure is called 'accommodation'. Between them, assimilation and accommodation are held to explain the gradual development of sophisticated conceptual structures which are used for modelling and for coping with the real world.

To simulate this process, the recogniser takes an input descriptor string and searches the template list for a good match. To keep it simple, a mean distance measure is used and only strings of the same length are considered as potential matches. If a match is made and the match is correct, then the template assimilates the input string by averaging the template with the input string. If no match is found, the system accommodates and a new template is formed using the input descriptor and the name of the word spoken. If a match is made but the match is incorrect, then accommodation is again called for and a new template is generated.

PROGRAM DETAILS

The zero-cross counts for the stimulus, $x(t)$, are sampled at a rate, r, and are transformed to lie on a scale between 0 and 50 (Fig 1A). Smoothing is effected using rectangular windows of width, w (Fib 1B).

$$s(t + w/2) = \sum_{i=t}^{t+w-1} x(i)/w \qquad (1)$$

The second order differentiation of the smoothed function (Fig 1C) is achieved using two points at a distance b units on either side of the point in question

$$s(t) = -s(t - b) + 2s(t) - s(t + b) \qquad (2)$$

A segment border is deemed to occur when this function passes from being negative to greater than a positive threshold value, h, and vice versa. The first segment begins with the onset of the utterance.

A word is composed of p segments and each segment is assigned a single label which is the value of s(t) at the midpoint of the segment. As a consequence, a short word would typically result in a description composed of a string of 2 - 6 labels, l(i). An input description is compared to all templates in store and the average distance is computed

$$d = \sum |l(i,s) - l(i,k)|/p \qquad (3)$$

where l(i,s) is the ith label in the input descriptor
l(i,k) is the ith label in the kth template of length p segments.

The template giving the lowest average distance value is used as the best match. If this is a correct guess and if d is lower than a criterion value, D, then the input descriptor is blended with the template

$$l'(i,k) = [l(i,s) + f.l(i,k)]/(f+1) \qquad (4)$$

where f is the number of input descriptors previously blended to make this template. Otherwise, a new template is created.

The system clearly requires a number of parameters to be optimised. The functioning of the system can be illustrated by a working configuration and a simple test. First, the following parameters were fixed on the basis of convenience and informal exploration; r = 50Hz, D = 2.5, h = 2. The parameters w and b were then both systematically varied between 1 and 9 while noting the percentage correct matches. A list of 13 isolated words (digits 1-9, oh, zero, plus, minus) were read aloud 17 times by a single speaker and the zero cross values stored on disc to be processed by the system for each set of parameters. The best performance (86% correct for the last 7 trials) was obtained for w = 5 and b = 3 (i.e. a smoothing window of 100 msec and a differentiation width, Δt, of 60 msec). The system began with no templates and finished with 59 templates after 221 stimuli. N.B. The Figures are based on a finer grain analysis with r = 200Hz, w = 50 msec., and b = 30 msec.

DISCUSSION

Because many words of input are matched to existing templates, the assimilation process keeps the number of templates to a minimum. Clearly different pronunciations of the same word will, however, automatically be allocated a separate template. As a consequence, garbage descriptor strings

will rarely be able to corrupt an existing template. Of course, the distance criterion for matching is a key parameter of the system. If this is too demanding then few matches will be found and new templates will proliferate. Similarly, the weighting of the averaging function used in the assimilation process will influence the stability of the system.

Because the input descriptors in this demonstration are very simple, they do not effectively discriminate between many acoustically similar words. Nevertheless, the system works moderately well (approximately 85% for a single speaker) after training on a modest vocabulary of twenty words with a mixture of similar and dissimilar words drawn from the Acorn Speech Chip vocabulary. The vocabulary was specifically chosen to illustrate the strengths and the weaknesses of the system during demonstrations. The test module works by accepting a single spoken word and attempts to make a match from existing templates. The user is then required to confirm or deny the system's suggestion before the accommodation or the assimilation routine is invoked. The system grows naturally by repeating this cycle.

The proposal identifies the program's template/word label pair with Piagetian schemata. Normally, such schemata would be thought of by psychologists as much richer structures. The current schemata merely combine auditory analysis (descriptor list) with appropriate responding (uttering the word label). Nor has any attempt been made to organise schemata either hierarchically nor heterarchically. Even in this limited context, this might be done by associating units with similar input descriptors, same word labels or word labels with similar meanings. We might even link units which are frequently sequentially associated. Such systems would take us a long way from the 'talking typewriter' ideal of most speech recognition endeavours by generating response errors such as synonyms and false anticipations., They would, however, take us a step further into the future, towards machines which make sense of the acoustic input in a manner analogous to human appreciation of speech.

REFERENCES

[1] R. Meddis, 'Exploiting physiological codes in ASR', this volume, (1986).
[2] D. Marr, 'Vision', Freeman and Company, (1982).
[3] J. Piaget, 'The origin of intelligence in the child', Penguin, (1977).
[4] R. Meddis, 'Simulation of mechanical to neural transduction in the auditory receptor', J.A.S.A., Vol. 79, 702-711, (1985).

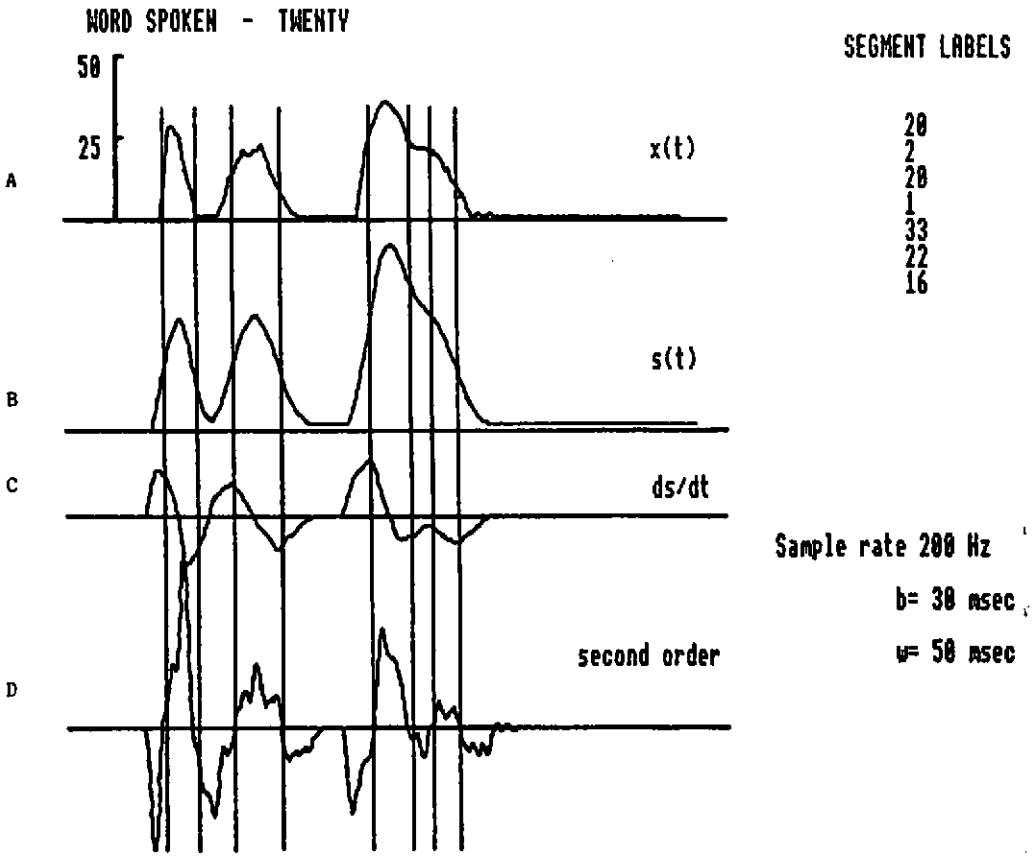

Fig. 1. Derivation of segments and their labels from zero-cross function. A. zero-cross count; B. smoothed function; C. first order differentiation; D. second order differentiation. See text for explanation.

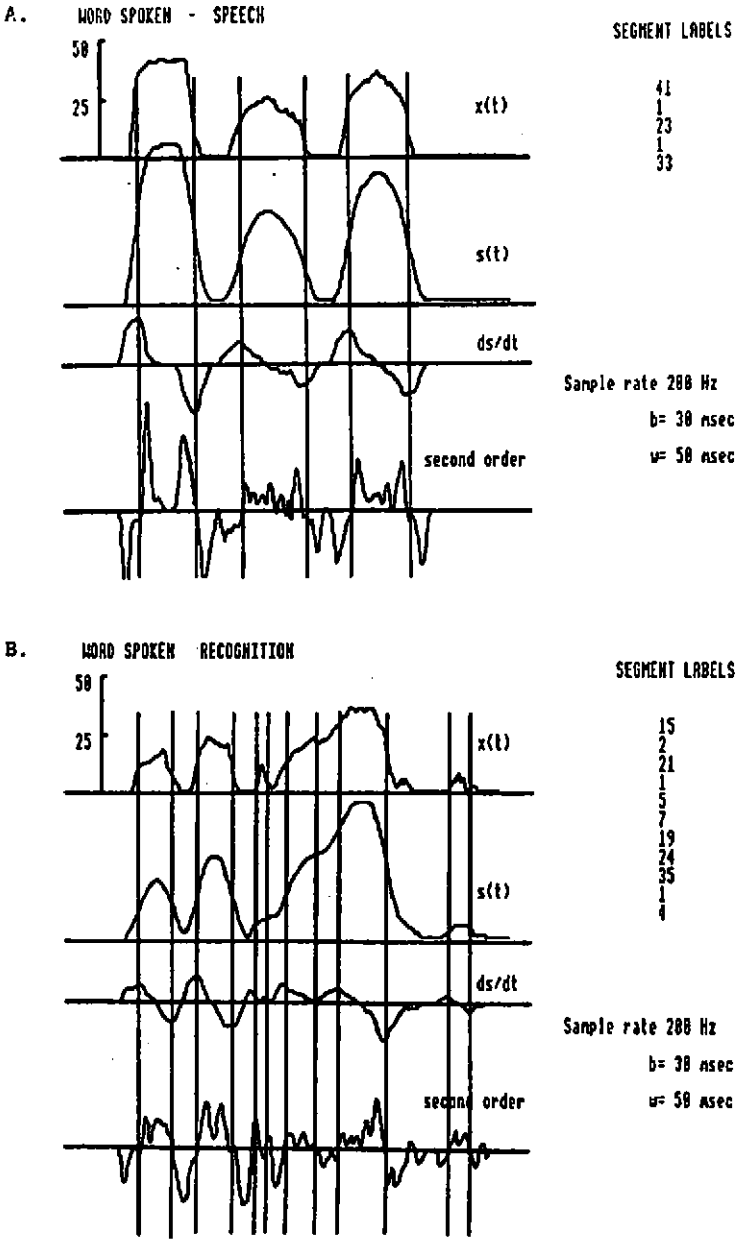

Fig. 2 A-D. Four examples of segmentation and labelling procedure using the words A. 'speech'; B. 'recognition'; C. 'leak'; D. 'new'. See text.

Fig. 2 continued.