INITIAL SPEECH SOUND PROCESSING IN SPOKEN WORD RECOGNITION

Phillip Dermody, Kerrie Mackie and Richard Katsch

Speech Communication Research Section
National Acoustic Laboratories

ABSTRACT - The present study uses the gating paradigm to investigate initial speech sound (ISS) processing in spoken word recognition. Results are presented for spoken words and consonant-vowel (CV) syllables, which both show very early recognition of the ISS. Acoustic analyses of the ISS show similarities between the words and syllables and are consistent with the templates proposed by Stevens and Blumstein (1978). It is suggested that the time course of ISS perception indicates the need to change present models of spoken word recognition.

INTRODUCTION

The failure to integrate the findings of speech perception research based on acoustic-phonetic studies and the results of cognitive investigations of spoken word recognition has been raised as a significant issue (Marslen-Wilson, 1983;Pisoni et al,1985). There is an obvious need to integrate information in these two areas of research. Models of spoken word recognition must be constrained by data about acoustic-phonetic processing and results from speech perception studies need to be validated in terms of speech recognition performance.

One possible focus for consideration of the interaction between speech perception research and spoken word recognition is the initial speech segment (ISS). The ISS is the beginning portion of the spoken word which is used to begin processing the word. Speech perception research has produced a large body of data about the acoustic and phonetic structure of initial speech sounds, especially for stop consonants. The ISS is also a critical determiner of how word candidates might be selected from the auditory lexicon. The present study investigates the ISS for words and syllables and considers these in terms of the temporal course of spoken word recognition proposed by current word recognition models.

ISS IN SPOKEN WORDS

In our first experiment we have replicated a study by Grossjean (1980) in which spoken words are presented in increasing durations (or gates) to listeners for identification. On the first trial the listener hears the first 30 milliseconds (ms) of the word (gate 1) and on the second trial hears the first 60 ms of the word (gate 2). The gates increase in length until the whole word is presented. The listener's task is to guess/recognise the word on each trial. In our study this gating paradigm was modified to focus on the processing of the ISS. That is,the listeners were instructed to identify the first sound in the word even if they could not attempt to guess the whole word. The instructions also asked the listeners to guess the word as soon as possible in the gating sequence. The gate at which the listener reliably identifies the sound or the word and then consistently repeats this response for successive gates is called the isolation point.

The stimulus words ranged from consonant-vowel-consonant words to words of three syllables. These words were recorded by a trained male speaker onto a computer based speech storage/editing system. The words were digitised at a sampling rate of 36KHz. Each word was then visually displayed and the beginning and endpoints of the word were marked. The gates of the words were

then marked by moving the end marker. The endpoint for the first gate was 30 ms from the beginning marker, for the second gate 60 ms from the beginning and subsequent gates sucessively marked using increments of 30 ms until the end of the word. The endpoints for each gate were always moved to the nearest zero crossing to avoid clicks. The gated words were then output to tape in sequential order with a 4 second inter-stimulus interval.

Subjects were undergraduate university students, aged 20 to 30 years, with normal hearing and were native English speakers. Testing was carried out in a quiet listening room with subjects wearing TDH49 earphones in sound attenuating circumaural headsets. Responses were recorded by the subjects on answer sheets.

Figures 1 and 2 show some typical results of these studies. Figure 1 shows the isolation point per gate duration for the first sound, second sound and for the whole word "DUCK" for 25 listeners. The vertical axis indicates the percentage of subjects who obtained an isolation point at each gate duration. From these data it can be seen that the first phoneme is recognised with a high degree of accuracy in the first 30 ms of the word. There is some overlap between the recognition of the consonant and the vowel and considerable temporal variability between the subjects for the point of identification of the full word.

The results for the word "PARTICLE" are presented in Figure 2. In this case there is again very good identification of the first phoneme in the first 30 ms. There is however, less overlap between the recognition of the vowel, and more variability in the identification of the final word. Figure 2 also shows the isolation point for the first possible word (ie."PART").

ISS IN CV SYLLABLES

From these gating data on spoken word recognition we can conclude that listeners have good recognition of the ISS in the first 30 ms of a word. In order to further explore how rapidly listeners can identify the ISS we carried out a further study in which we examined the identification of the initial consonant in CV syllables. In this study the stimulus set was the six stop consonants plus the vowel /a/ spoken by the same speaker as for the words. These CV syllables were gated in a similar manner to the words with the gate duration reduced to 10 ms. The listeners were given a closed set of six response alternatives and asked to identify each stimulus. In the word gating paradigm the incremental gates were presented sequentially throughout the words. For the CV syllables the different gate durations of all stimuli were presented in random order, with four repetitions of each gate duration being presented.

Figure 3 shows the results of 20 subjects averaged over the four repetitions of each gate duration from 10 to 150 ms in 10 ms increments. The results indicate that the overall recognition performance for gates from 20 ms duration and above is very high, while the recognition at the 10 ms gate is above chance performance.

There are two aspects of these data that are of interest to us. First,we considered what acoustic information might be used to correctly identify the ISS in stop consonants within the first 10 to 20 ms and second, how the ISS might be used in spoken word recognition. To investigate the question about acoustic information, we compared the acoustic analysis of the ISS with the templates suggested by Stevens & Blumstein (1978).

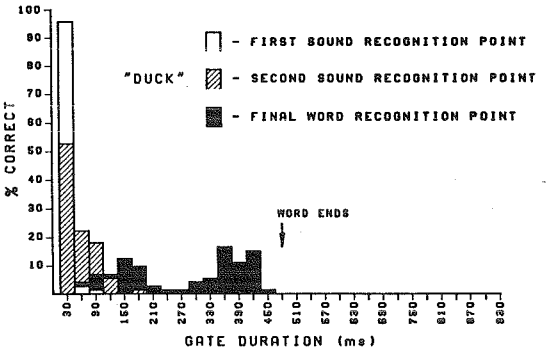

Figure 1. Data for incremental 30ms gates. First sound is /d/,second sound is /u/ and word is /duk/.

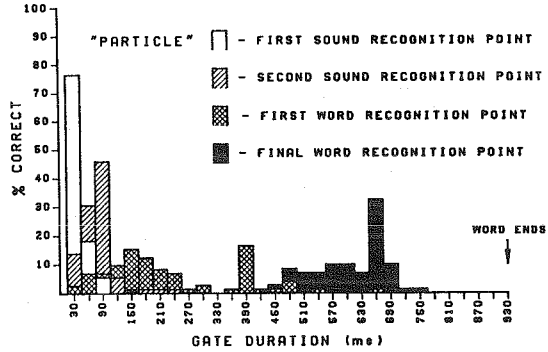

Figure 2. Data for incremental 30ms gates. First sound is /p/, second sound is /pa/. First word is /pat/,second word is /patIkəl/.

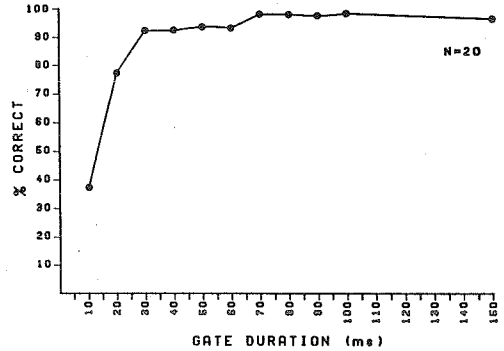

Figure 3. Average percent correct identification of the stop consonants plus /a/ for each gate duration.

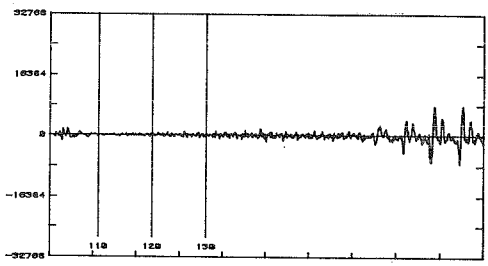

Figure 4. Waveform analysis of the first 80ms of /patIkəl/ showing the points taken at 10,20 and 30ms durations.

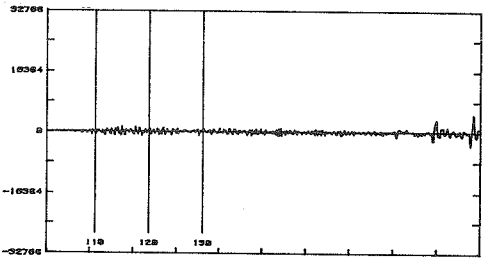

Figure 5. Waveform analysis of the first 80ms of /pa/ showing the points taken at 10,20 and 30ms durations.

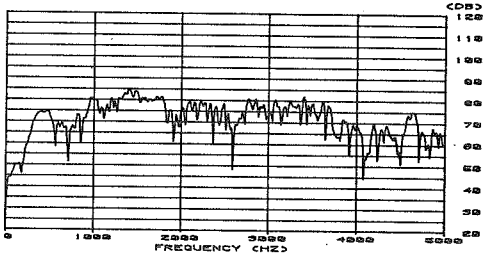

Figure 6. FFT analysis of the first 30ms of /patIkəl/.

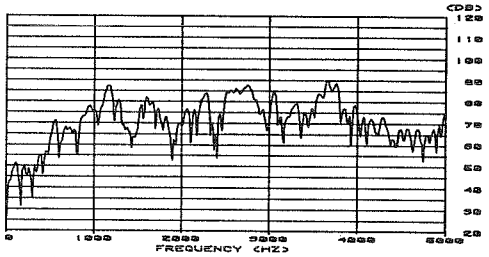

Figure 7. FFT analysis of the first 30ms of /pa/.

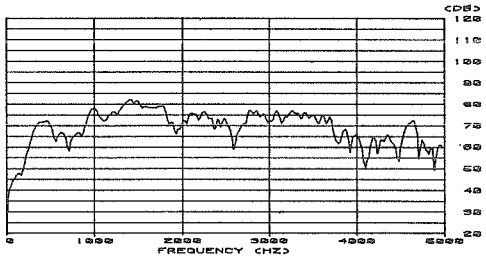

Figure 8. FFT analysis of the first 20ms of /patIkəl/.

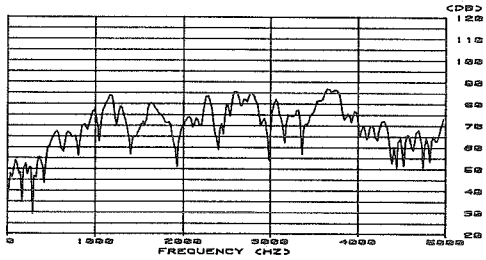

Figure 9. FFT analysis of the first 20ms of /pa/.

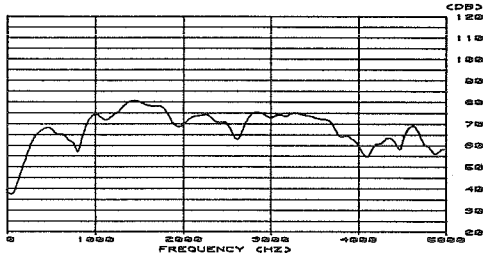

Figure 10. FFT analysis of the first 10ms of /patIkəl/.

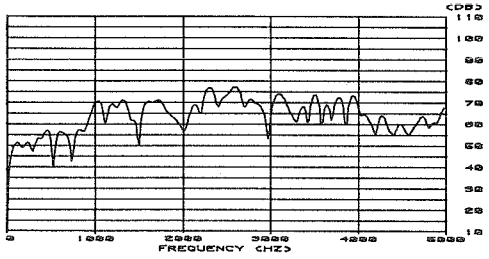

Figure 11. FFT analysis of the first 10ms of /pa/.

ACOUSTIC ANALYSIS

Figures 4 & 5 show the waveform for the first 80 ms of /pa/ and /patIkəl/. The first 10, 20 and 30 ms are marked to show the analysis periods that were used in this study. Figures 6 & 7 compare the first 30 ms of the word /patIkəl/ to the first 30 ms of /pa/ from the CV syllable set. An FFT analysis (ILS v.5.0) was carried out on the speech which was resampled from the test tapes at 10K samples per second. The resulting spectra (figures 6 & 7) conform to the falling-flat spectrum described by Stevens & Blumstein (1978) as characterising labial plosives.

Figures 8 & 9 present the same analysis for the two stimuli using the 20 ms gate and figures 10 & 11 show the analysis for the 10 ms gates. While these are not identical in detail, in all cases the results are consistent with the template for the labial plosive posed by Stevens & Blumstein. In figures 10 & 11 the frequency resolution is reduced,due to the shorter duration analysed, ie. due to the fewer sampled points used in computing the spectrum. These data are consistent with the view that there is a characteristic acoustic pattern which some listeners may utilise at durations of 10 ms to begin processing the /p/ in both syllables and words.

TEMPORAL COURSE IN SPOKEN WORD RECOGNITION

Studies in the area of word recognition suggest that the recognition time for words is typically about 200ms (Marslen-Wilson, 1983) and that this is considered fast for words in which the overall duration may be about 400 ms. The present data suggest that for some listeners, the word candidates from the auditory lexicon may start to become available at around 10 ms or less. This would certainly provide time for the elaborate search procedures for word candidates envisaged in present word recognition models. However, it also suggests the possibility of alternate explanations of word recognition based on fast processing of the elements of speech sounds which may operate while the speech recogniser is waiting for additional low frequency spectral information to accumulate.

The model of a fast processing system of speech sound elements while acoustic information is accumulating to provide a spectral representation of diphones, syllables or words is consistent with at least one model of speech processing and lexical access (Klatt, 1979). We are at present investigating the implications of these observations in further studies of the use of the ISS in spoken word recognition.

REFERENCES

GROSSJEAN,F.(1980)"Spoken word recognition processes and the gating paradigm", Perception & Psychophysics, 28,267-283.

KLATT,D.(1979)"Speech perception- a model of acoustic-phonetic analysis and lexical access", Journal of Phonetics,2,279.

MARSLEN-WILSON,W.(1983)"Perceiving speech and perceiving words", Abstracts of the Tenth International Congress of Acoustics,IIA,39-42.

PISONI,D.,NUSBAUM,H.,LUCE,P.,& SLOWIACZEK,L.(1985)"Speech perception, word recognition and the structure of the lexicon",Speech Communication,4,75-95.

STEVENS,K.,& BLUMSTEIN,S.(1978)"Invariant cues for place of articulation in stop consonants",Journal of Acoustical Society of America,64,1358-1368.