A STUDY OF FREQUENCY TRANSITION DETECTION USING SYNTHETIC VOWEL SOUNDS

P. Cosgrove and J.P. Wilson

Department of Communication and Neuroscience,
University of Keele, Staffordshire, ST5 5BG.

INTRODUCTION

Frequency transition detection is an important area for investigation in the field of speech perception as formant frequency changes function as major cues for the recognition of many speech sounds. There have been many previous experiments in speech-related research using stimuli containing frequency changes; but earlier studies suffered from the problem that the stimuli used were not "speech-like" in structure, and therefore did not sound like speech (e.g. Brady et al, 1961). In recent years however, this situation has been remedied, but there have been few systematic studies of detection thresholds for frequency changes in speech.

The work reported here represents a preliminary study of formant frequency transition detection thresholds for certain vowel phonemes, and forms part of a wider-based project whose aim is to investigate the psychoacoustical constraints on speech recognition.

METHODS

Generation of stimuli. The stimuli were generated, off-line, by an Apple IIe microcomputer using a software parallel-formant speech synthesizer. The synthesis program is a software implementation of the JSRU parallel-formant speech synthesizer (Rye and Holmes, 1982). The stimulus parameters were constructed from tables published in Holmes et al (1964) and Ainsworth (1974).

All the stimuli used were isolated vowels (or isolated formants) of 200 ms duration. The vowels were /3/, /i/ and /a/, chosen to give a wide representation of F_1-F_2 space with a minimal sample of stimuli (see Fig. 3). The formant frequencies chosen for the above three vowels were as follows :- /3/ - F_1=580 Hz, F_2=1420 Hz, F_3=2620 Hz; /i/ - F_1=250 Hz, F_2=2320 Hz, F_3=2740 Hz; /a/ - F_1=790 Hz, F_2=1060 Hz, F_3=2500 Hz. All the vowels had a fixed F_4=3500 Hz, and a fixed F_0=100 Hz. The bandwidths were 80, 100 and 120 Hz for F_1, F_2 and F_3 respectively.

Transition durations of 10, 20, 40 and 80 ms were chosen to be typical of the range of frequency transitions in speech (Nábelek and Hirsh, 1969; Tsumura et al, 1973). The formant transitions all occurred at the end of the stimulus in order to minimise pitch cues from steady-state segments. Except where otherwise stated, all transitions were upward.

General psychophysical considerations suggested that a logarithmic scale in which equal divisions represent equal ratios of rate-of-change of frequency would be more appropriate than a linear scale. The interval ratio chosen was

√2. Greater ratios tended to produce an "all or nothing" situation in which listeners would score 100% correct on one stimulus level and little better than chance on the next. Smaller ratios required too much adjustment to the abilities of individual subjects, although sometimes a smaller ratio was included. A total of six different levels of stimulus difficulty were used in each experiment.

All the stimuli were generated using a Rosenberg glottal pulse shape in the synthesis program. This was employed because it is a good approximation to a human glottal pulse shape: it produces reasonably "natural" sounding synthetic speech. The stimuli were filtered with a passband of 100-5000 Hz.

Procedure. A two-alternative forced choice procedure was employed for the experiments. Subjects were seated in a sound-proofed booth and the stimuli were presented binaurally via Sennheiser HD414 headphones from a Digital-to-Analogue converter connected to the computer. Visual feedback of errors was given. Each stimulus pair contained a steady-state and a transition stimulus, the task for the subject being to identify which of the two intervals contained the transition. Each subject was allowed a few trial runs until consistent performance was achieved, but no attempt was made to overtrain.

The stimuli were presented at an overall sound level of 85 dB SPL. This proved to be an optimal and comfortable listening level, and ensured that all formant amplitudes were above threshold although not necessarily equal. Since this inequality occurs in natural speech this was judged to be acceptable.

Analysis of data. Formant frequency transition detection thresholds were obtained at the 75% correct level. Each subject responded to 100 stimulus pairs at each of the 6 different levels presented in pseudorandom order. Raw data were fed into a program designed to fit a logistic function to binary "stimulus-response" data (Foster, 1986). T-tests were then performed on both individual and pooled data. Comparisons were made between the detectability of four-formant and isolated formant conditions, and between the first and second formants of a particular vowel. In cases where a subject had completed trials on the same formant for two different vowels, a t-test on the differences between the two was performed.

RESULTS AND DISCUSSION

As expected, all conditions show a decreasing threshold with transition duration with greatest changes occurring between 10 and 20 ms (Fig. 1).

Fig. 1(a) shows transition detection thresholds of the F_2 of /3/ under isolated and four-formant conditions. Predictably, the F_2 transition was slightly but significantly easier to detect when isolated than within the four-formant context ($p<0.05$, 1-tailed). The differences for /i/ were more marked than for

Fig. 1. Formant frequency transition thresholds (%) plotted as a function of transition duration for the conditions indicated.

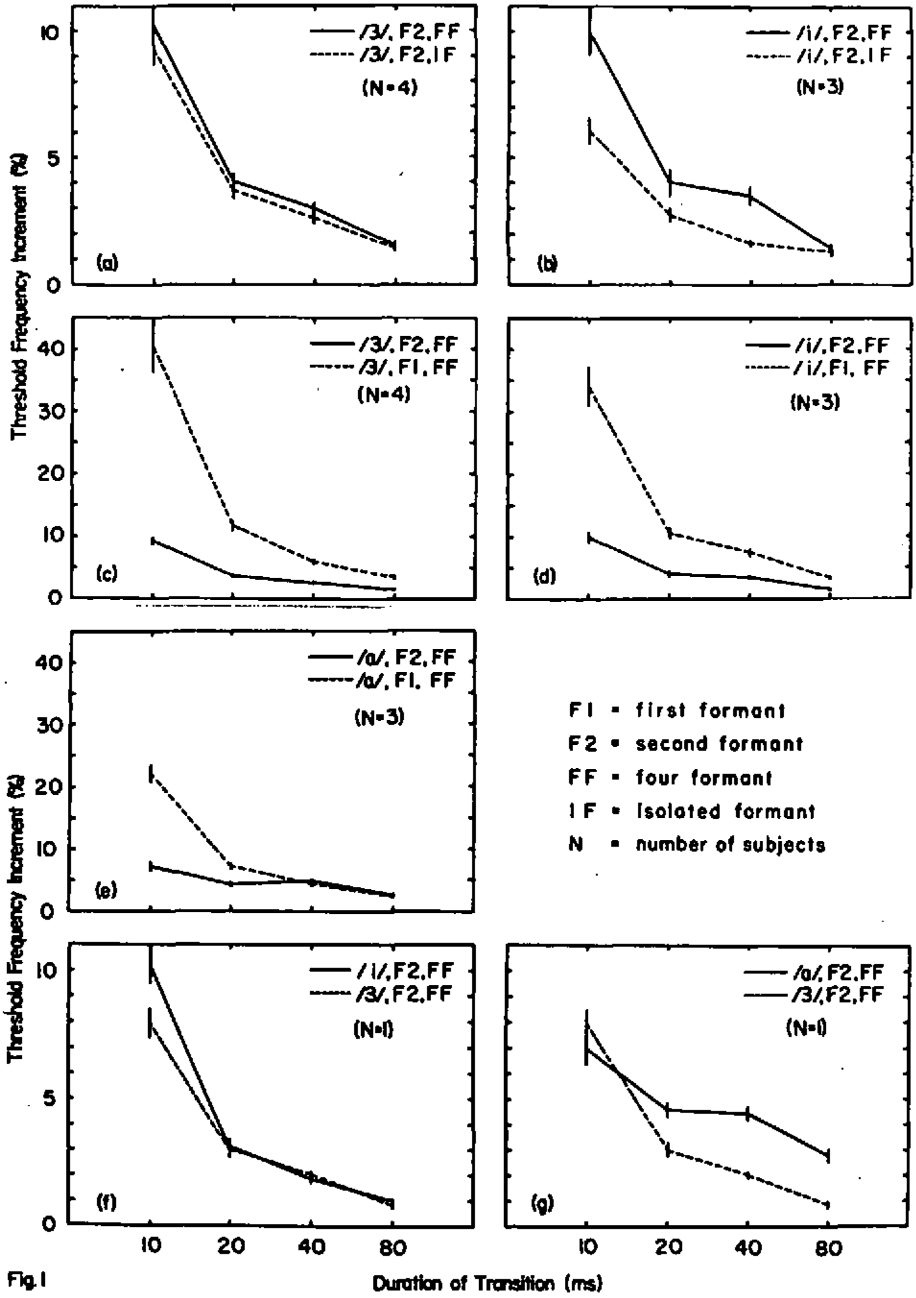
Threshold Frequency Increment (%)
—— /3/, F2, FF
------ /3/, F2, IF
(N=4)
(a)
—— /i/, F2, FF
------ /i/, F2, IF
(N=3)
(b)
—— /3/, F2, FF
------ /3/, F1, FF
(N=4)
(c)
—— /i/, F2, FF
------ /i/, F1, FF
(N=3)
(d)
—— /a/, F2, FF
------ /a/, F1, FF
(N=3)
(e)
F1 = first formant
F2 = second formant
FF = four formant
IF = Isolated formant
N = number of subjects
—— /i/, F2, FF
------ /3/, F2, FF
(N=1)
(f)
—— /a/, F2, FF
—— /3/, F2, FF
(N=1)
(g)
0
5
10
10
20
30
40
10
20
40
80
Duration of Transition (ms)

Fig.1

/3/ ($p<0.001$, 1-tailed; see Fig. 1(b)), possibly due to a downward masking effect of F_3 upon F_2 (see Fig. 3 /i/). At detection threshold F_2 and F_3 were within one critical band (Scharf, 1970) for /i/ but much more widely separated for /3/.

The isolated formant in these experiments is not a sinewave or noise band, but a formant with glottal pulse excitation at F_0. It is not particularly "speech-like", but serves as the most appropriate comparison for the possible influence of other formants.

The transition thresholds for F_1 detection were compared with those for F_2 in all the three vowels tested. From fig. 1(c-e) one can see that the differences are quite clear, requiring about twice as much change in F_1 for detection. For /3/, /i/ and /a/ the differences were significant ($p<0.001$, 1 or 2-tailed) that the F_2 transition was easier to detect than the F_1.

Because of time constraints individual subjects tended to be tested on one particular vowel. With different subjects direct comparison between vowels becomes impossible. However, one subject (PC) performed all of the experiments reported in order that some direct comparisons could be made.

In fig. 1(f) the detection thresholds of the F_2's of /3/ and /i/ are plotted. For one subject it was found that the F_2 transitions of /3/ were easier to detect than those of the vowel /i/ ($p<0.05$, 1-tailed). This was in line with the results shown in fig. 1(a) and (b) where the F_3 of /i/ appeared to interfere with the detection thresholds of F_2. The vowel /3/ has relatively widely (and evenly) spaced formants in comparison to /i/.

Fig. 1(g) illustrates that the F_2 transitions of /3/ are easier to detect than those of /a/ ($p<0.001$, 1-tailed). For /a/ the F_1 and F_2 are rather close to each other, and one would possibly expect an upward masking effect on the F_2. Even if no prediction about direction had been possible in this particular case, the result was still significant ($p<0.01$, 2-tailed).

One subject (PC) completed trials on both upward and downward transitions of the F_2 of /3/ (not illustrated). Gardner and Wilson (1979) found that at 1 kHz the threshold frequency change for 62.5 ms pure-tone upsweeps was lower (0.85%) than for downsweeps (1.5%). Collins and Cullen (1984), using a noise masking technique, also found lower thresholds for upsweeps. In the present study (at F_2=1420 Hz) downsweeps had a slightly lower threshold for the three shorter durations, and slightly higher at 80 ms. Overall, however there was no significant difference. Interestingly, the mean for the values (40 and 80 ms) around 62.5 ms and up and downsweeps (1.42%) agrees very closely with Gardner and Wilson's mean (1.18%).

As the threshold values in fig. 1 all appeared to decrease consistently as transition duration increased, the data were replotted as threshold (%) x duration (ms) as a function of transition duration. These are shown in fig. 2. On average (fig. 2(h)) this relationship appears to be independent of duration. The implication of this is that sensory integration is taking place so that threshold formant frequency increment is inversely proportional to transition duration. It should be noted that this is the inverse of rate-of-change

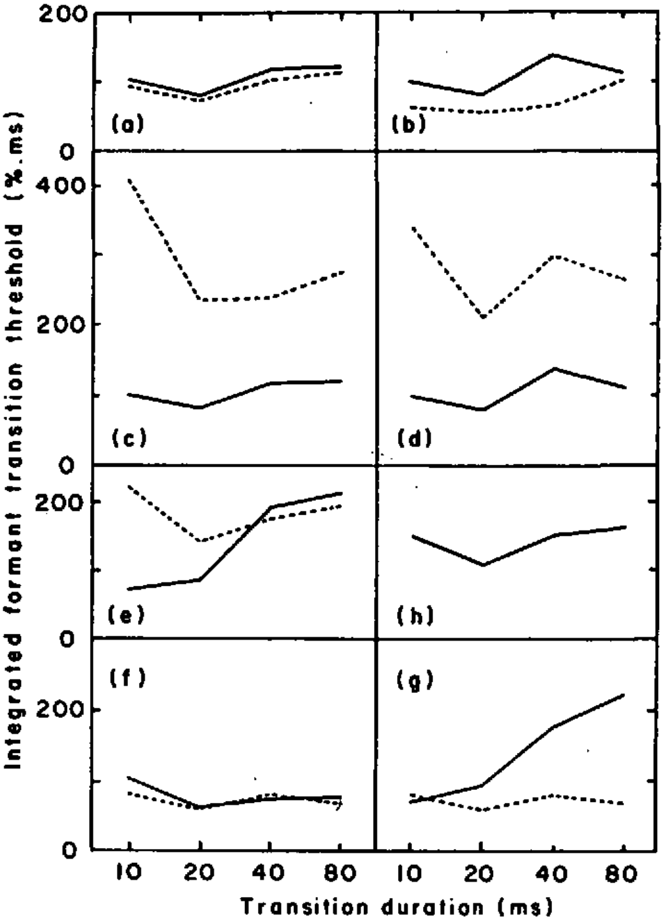

Fig. 2. Integrated formant transition thresholds (%.ms) plotted as a function of transition duration. These were obtained by multiplying the values in fig. 1 by the stimulus durations concerned. The conditions (a-g) are the same as fig. 1(a-g) and (h) represents the average for the four-formant stimuli.

detection where threshold would be proportional to transition duration.

The integrative relationship appears to hold over the range 10-80 ms which covers (almost completely) the durations found in speech transitions. It appears likely, however, that it must break down for very short durations where the frequency excursion would become very large, and for longer durations than 80 ms where the integrating time constant of the auditory system would be exceeded. One practical implication of this is that related experiments may need only be performed at a single duration within the 10-80 ms range, and the results can then be compared with other experiments where a different duration has been used.

In an attempt to obtain further generalisation of the results, the mean values (threshold (%) x duration (ms)) from fig. 2 were plotted against frequency on a critical band scale (Barks). The threshold values (% x ms) appeared to decrease systematically with critical band position of the formant concerned. If, however, threshold frequency increment x duration (Hz x ms) was plotted versus critical band position, the opposite trend was noted. It was decided, therefore, to express the threshold frequency increments in critical bands (x duration). These are plotted in fig. 3 for the three vowel phonemes /a/, /3/ and /i/, and in a combined figure (lower).

There is now no overall trend with frequency, and the four-formant values all lie within a range of 2:1. The general implication of this is that when the threshold frequency increment is expressed in critical bands it is independent of frequency position and formant number. Thus the differences analysed above would appear to be due chiefly to frequency position. The mean value of this threshold across all four-formant stimuli is 8.9 critical band x ms. If one takes a 250 ms integrating time for the auditory system this indicates a possible asymptotic threshold of 0.036 critical bands, e.g. 5.7 Hz at 1 kHz. This can be compared with the pure-tone frequency difference limen which ranges from 1 to 3 Hz at 1 kHz according to different authors and procedures. A more direct comparison, mentioned above, is with the results of Gardner and Wilson (1979). The mean values for upsweeps (U) and downsweeps (D) were 3.3 and 5.9 Bark.ms respectively. These are shown by the dotted line (fig. 3, lower). It appears from this that changes within a speech-like sound are about twice as difficult to detect as within a pure tone psychophysical context.

CONCLUSION

The results reported here indicate that frequency transition detection threshold is inversely proportional to transition duration over 10-80 ms and therefore lends no support for the existence of specific rate-of-change detectors.

Threshold is approximately constant over frequency when expressed in critical bandwidths (Barks) at that frequency, and therefore lends no support for the existence of specific formant detectors. In percentage frequency terms, however, F_2 does appear to have a lower threshold than F_1.

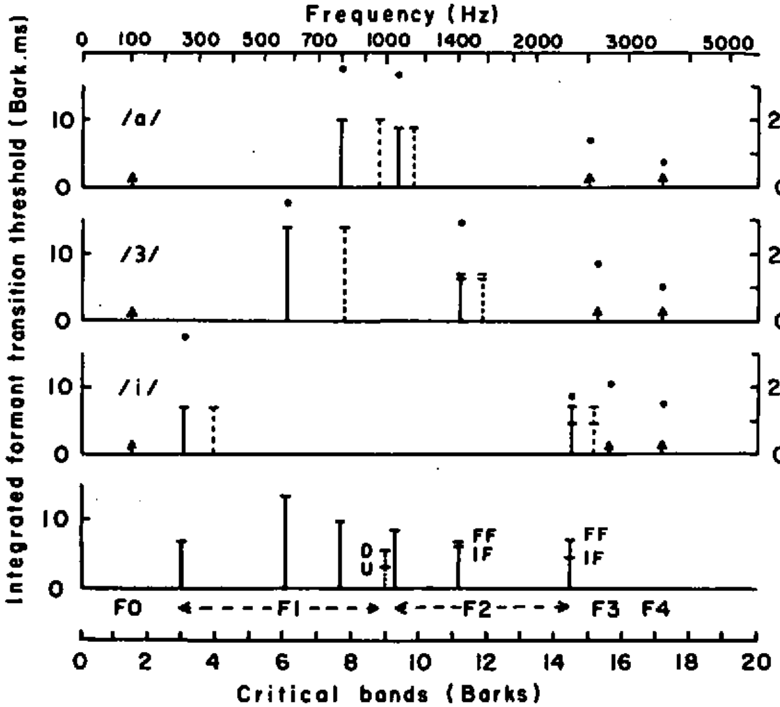

Fig. 3. The solid vertical bars represent the integrated threshold values expressed in Bark.ms, obtained by averaging across duration of the relevant functions shown in fig. 2. The three vowel sounds /3/, /a/ and /i/ are represented separately in the three upper diagrams and together, below. They are plotted as a function of frequency (indicated above), spaced on a uniform critical band scale (Barks) to indicate spacings where masking might occur. The dots and right-hand scales indicate the relative levels of the various formants. The positions of F_0, F_3 and F_4 are indicated by the upward arrowheads. The dashed bars in the upper three diagrams indicate the highest upward shifts which occurred at the 10 ms duration. The lower horizontal bars on some lines represent the isolated formant (IF) conditions. The dotted line on the lowest diagram represents the data of Gardner and Wilson (1979) normalised in the same way, for downward (D) and upward (U) tone sweeps.

Perhaps surprisingly, the detection of a transition within one formant of a four-formant complex is only about twice as difficult as the tasks of frequency discrimination and the detection of a pure-tone frequency sweep. Clearly this should be tested in a more direct comparison.

ACKNOWLEDGEMENTS

This work was supported by the Speech Research Unit, Malvern. We would like to thank Dr S.R. Pratt for helpful suggestions on the manuscript.

REFERENCES

W.A. Ainsworth, 'Performance of a speech synthesis system', Int. J. Man-Machine Studies., Vol. 6, 493-511, (1974).

P.T. Brady, A.S. House, and K.N. Stevens, 'Perception of sounds characterized by a rapidly changing resonant frequency', J.A.S.A., Vol. 33, 1357-1362, (1961).

M.J. Collins and J.K. Cullen, Jr, 'Effects of background noise level on detection of tone glides', J.A.S.A., Vol. 76, 1696-1698, (1984).

D.H. Foster, 'Estimating the variance of a critical stimulus level from sensory performance data', Biol. Cybern. Vol. 53, 189-194, (1986).

R.B. Gardner and J.P. Wilson, 'Evidence for direction-specific channels in the processing of frequency modulation', J.A.S.A., Vol. 66, 704-709, (1979).

J.N. Holmes, I.G. Mattingly, and J.N. Shearme, 'Speech synthesis by rule', Language and Speech, Vol. 7, 127-143, (1964).

I. Nåbelek and I.J. Hirsh, 'On the discrimination of frequency transitions', J.A.S.A., Vol. 45, 1510-1519, (1969).

J.M. Rye and J.N. Holmes, 'A versatile software parallel-formant speech synthesizer', J.S.R.U. Research Report No. 1016, (1982).

B. Scharf, 'Critical Bands'. In Foundations of Modern Auditory Theory (ed. J.V. Tobias), New York: Academic Press, 157-202, (1970).

T. Tsumura, T. Sone, and T. Nimura, 'Auditory detection of a frequency transition', J.A.S.A., Vol. 53, 17-25, (1973).