STUDY OF THE ROLE OF FORMANT TRANSITIONS IN VOWEL RECOGNITION USING THE MATCHING PARADIGM

L.C.W. Pols, G.W. Boxelaar, and F.J. Koopmans-van Beinum

Institute of Phonetic Sciences, University of Amsterdam,
Herengracht 338, 1016 CG Amsterdam, The Netherlands

ABSTRACT

In a well-known paper by Lindblom and Studdert-Kennedy [11] the role of formant transitions in vowel recognition was studied by using continua of synthetic /wVw/, /jVj/ and /#V#/ stimuli which had to be identified as either /i/ or /u/. They concluded that in the recognition of monosyllabic nonsense speech the identity of a vowel is determined not solely by the formant-frequency pattern at the point of closest approach to target, but also by the direction and rate of adjacent formant transitions.

As an extension of earlier identification experiments with vowel segments isolated from natural conversational speech, with and without the transitions to neighbouring consonants, we have studied the role of formant transitions in more detail. We asked subjects to match stationary vowel sounds with dynamic stimuli consisting of various formant transitions. Of these stationary sounds one formant frequency could be controlled by the subject. The dynamic stimuli were of the type /V_1V_2/ or /$V_1V_2V_1$/, without /V_i/ necessarily being a representative Dutch vowel, and the subjects were asked to match /V_2/. A great many possible transitions in the stimuli had to excluded because they elicited a perception of a consonantal /w/ or /j/, which could not be matched with a stationary vowel sound. So far there is no indication of strong overshoot effects. Preliminary data suggest that the matching results could be influenced by the presence or absence of a vowel boundary near the target value.

INTRODUCTION

There are many physical differences between a stationary synthetic vowel sound presented in isolation and a similar vowel in conversational speech. For a listener it is nevertheless not too difficult to label both signals with the same vowel symbol. In order to study the process of vowel perception, one could try to evaluate the various properties of the signal which contribute to vowel identification.

In conversational speech these properties are so manifold (e.g. language, talker, acoustic environment, speech style, rate, sentence context, word structure, accent, prosody) that it is very difficult to vary them systematically. In the case of an isolated steady-state vocalic stimulus of lifelike duration and fundamental frequency, variation is confined to the frequency spectrum. We believe we know more or less what the relation is between spectral characteristics (e.g. formants or band filter representations) and vowel quality. One small step towards a slightly more natural vowel sound is the introduction of spectral variation. In natural speech there is hardly ever a stretch of speech longer than say 30 ms which can be considered to be unchanging. So, the listener has constantly to cope with dynamically varying speech characteristics. Several studies suggest that a vowel surrounded by plosives is better identified than a vowel in isolation because of the dynamic specification [16]. This could mean that mutual interaction between neighbouring speech sounds (coarticulation) far from blurring or degrading phonetic identities, actually enhances information transmission and is a prerequisite for naturalness. In order to get some understanding of how this could work we studied vowel identification and vowel matching with dynamically varying vowel-like stimuli. In the preceding paper of this symposium [9] some data were presented on the identification of vowel segments excerpted from conversational speech. Two variables were specifically considered. One was the importance of adding to the vowel nucleus more and more transitional information towards preceding and following consonant, and the other was the effect of different response categories.

Although some very interesting results came out of these experiments, one of the drawbacks is that the listener could only specify his percept in terms of a limited number of predefined categories. If a specific /ɑ/-type stimulus is slightly more /ɔ/-like than another, this kind of detail could not be captured in listeners' responses. We figured that by using the matching paradigm we could elicit more finely-grained judgments. Most of the studies in this area reported in the literature have made use of one of two procedures: identification or matching.

SHORT LITERATURE SURVEY

Formant transitions

A very important paper about the role of formant transitions in vowel recognition was that of Lindblom and Studdert-Kennedy [11]. They used synthetic vowel stimuli, presented either in isolation or in a symmetrical environment of /w/ or /j/. The formant frequencies of the vowel nucleus varied in 20 steps from /u/-like to /i/-like. Subjects were asked to label the vowel as either /i/ or /u/. There was a strong context effect that shifted the cross-over point between /u/- and /i/-judgments toward the locus of the adjacent semivowel. The results suggest that the transitions were permitted to undershoot the target frequencies for the vowel. Since in natural speech also the formant trajectories almost invariably undershoot their target values, these perceptual results seem to indicate an efficient mechanism to compensate for the articulatory undershoot associated with vowel reduction. Recently Kuwabara [10] found similar results for /uVu/ and /eVe/ stimuli and showed, by using a dichotic fusion paradigm, that "the boundary-shift mechanism is not located in the peripheral system, but is more likely a process of the central brain function".

Fujisaki and Sekimoto [4] studied the effect of time-varying resonance frequencies in open syllables where a one-way transition elicits perception of two contiguous phonemes, either /V_1V_2/ or /CV/. The used F_1 or F_2 transitions were similar to those used in our experiments. However, they applied three versions of each transition by truncating at 70, 80, or 95% of the formant sweep. For comparison the listeners heard static stimuli with formant frequency values in a range of 120 Hz around the target frequency. They concluded that "subjects do not necessarily perceive the terminal frequency of a truncated transition, but are capable of approximately extrapolating the target frequency, not only in non-speech (one-formant) stimuli, but also in speech (five-formant) stimuli". There were individual differences for the effect of consonant context and transition rate. Our main objection to this experiment is the very limited frequency range over which the static reference stimuli were allowed to vary. This range was 120 Hz around the target frequency and did not even include the terminal frequencies of the 70 and 80% truncated dynamic stimuli, not to speak of possible lower (undershoot)values.

We have similar objections to the matching experiment of Kanamori and Kido [8]. They asked subjects to select from a set of 15 stationary 150-ms vowel stimuli the one which most resembled the vowel in CVC-, VC-, or CV-type dynamic stimuli of 60- to 140-ms duration. The results showed a clear tendency towards overshoot, however, the characteristics of the 15 possible stationary vowels hardly allowed any other

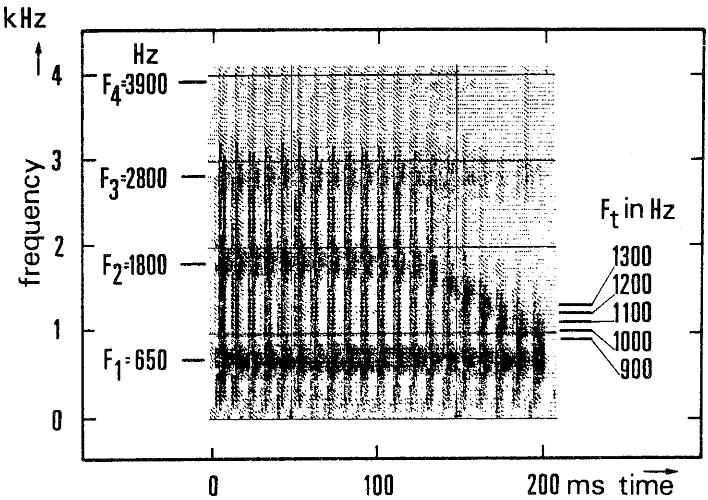

Fig.1. Digital spectrogram of one of the four-formant stimuli with F_2 falling from 1800 to 900 Hz. Other terminal frequencies are also indicated.

choice.

If one allows free control over the stationary vowel which has to be matched with a formant sweep, it is not at all clear that this will result in an extrapolation of that transition (overshoot). House and Neuburg [7] used formant sweeps going up or down over a range of 500 or 1000 Hz, with a duration of 100 or 150 ms. They consistently found, for all five subjects, an undershoot setting. Also for triangular sweeps of 150- or 200-ms duration and ranges of 180 to 530 Hz, the matching formant frequencies of the stationary vowels were located between locus frequency and final frequency. The results did not differ much for one-, two-, or three-formant stimuli.

Tone and band sweeps

Next we will mention a few psycho-physically oriented studies involving band sweeps and tone sweeps. Brady et al. [1] used pulse trains, with a repetition rate of 100/s, the tuning circuit had a bandwidth of 100 Hz, and the midfrequency was swept over 500 Hz up or down within 20 to 50 ms, the stationary signal preceding and following the sweep was also varied in duration. Subjects had to match the sweeping test stimulus with a stationary comparison signal (with adjustable resonant frequency) until the two signals were judged most alike. There was a strong tendency for the adjustments to be close to the terminal resonant frequency of the sweep. The longer the sweep duration and the longer the trailing end, the stronger this tendency.

Several papers have studied the detection and discrimination of tone sweeps [2,13] and band or formant sweeps. In these studies there are always two complicating factors. One is that in all non-symmetrical sweeps, especially those with a steady frequency before and after the sweep, the frequency-sweep discrimination could be based on frequency discrimination between initial and final frequency. The second is the interdependence of sweep height, sweep rate, and sweep duration. Because of this reason, Horst [6] used bands the center frequency of which swept symmetrically in time according to a bell shape. He found sweep-detection thresholds of 0.4 to 2% of the sweep height and sweep-discrimination thresholds of 1.0 to 3.5%. These values are considerable lower than the 4-9% reported by Danaher et al [3] and Mermelstein [12].

Mermelstein studied the difference limen (DL) for formant frequencies of steady-state vowels and dynamic vowels with symmetric formant transitions as in CVC-type stimuli. The DL for the time-varying CVC stimuli was significantly larger (60 Hz for F_1 and 176 Hz for F_2) than for steady-state vowels. On the basis of these results Mermelstein questions the perceptual significance of most of the F_2-shifts due to consonantal context as, for instance, observed by Stevens and House [15]. He also concludes that the improved identification for natural vowels in context over those in isolation [16], is not due to improved formant frequency discrimination.

Schouten [14] used tone sweeps of 20 to 50 ms duration, without steady state before or after the sweep, and rates of 0 to 60 oct/s. He found, with naive and minimally trained subjects that stimuli with zero or low sweep rates were judged to move 'down', whereas stimuli with high sweep rates received 'up' responses, irrespective of the actual direction. Discrimination between rising and falling tones did not differ significantly from discrimination between rising and level tones. This perceptual imbalance between upward and downward tone sweeps is found more often. Gardner and Wilson [5] propose evidence for direction-specific channels for the processing of frequency modulation. Selective adaptation experiments [17] suggest the existence of independent channels for amplitude and frequency modulation. Perhaps the selectivity to low-frequency modulations of 2-32 Hz could also be interpreted as a selectivity to upward or downward frequency sweeps.

It will be clear from this short and incomplete survey that many aspects of the detection, discrimination, identification, and matching of tone, band, and formant sweeps are still unknown. The experiment described below, situated halfway between psychophysics and speech perception, is one step towards a better understanding of this process.

EXPERIMENTAL PROCEDURE

We used the matching paradigm to investigate how formant transitions are interpreted by listeners. The stimuli were synthetic four-formant vowel-like signals. Three formants were kept constant and either F_1 or F_2 changed during the final 100 ms of the 200-ms stimuli. The fundamental frequency was continuously falling from 110 Hz to 90 Hz. Fig. 1 shows a digital spectrogram of one stimulus out of a set of five in which F_2 moves downward from 1800 Hz to 1300-900 Hz. With the given values for the other formants, this signal sounds like an /ε-œ/-type vowel which changes to an /ɑ/-type vowel. The formant transition has the form of half a sine wave. In order to prevent clicks the overall amplitude is gradually shaped over the initial and final 2 ms of the signal. By pressing a button the subject, seated in front of a keyboard with terminal, can listen to this signal over earphones as often as he wants. After 500 ms this signal is succeeded by the comparison signal, which is a stationary four-formant signal of 70-ms duration. By using a 'left' or 'right' button the frequency of one formant (F_2 for the stimuli in fig. 1) of the comparison signal goes up or down. This frequency is at the same time displayed on the screen as a position on a 32-point horizontal scale. The 32 logarithmically scaled formant values cover a substantial range around the terminal frequency of the changing formant in the test stimulus. Two successive positions along the scale are about half a jnd apart. The subjects never complained that they would have liked a larger, or more detailed, range.

The varying test stimulus was calculated beforehand by using digital filters and was generated by computer by reading the appropriate file of sampled data upon re-

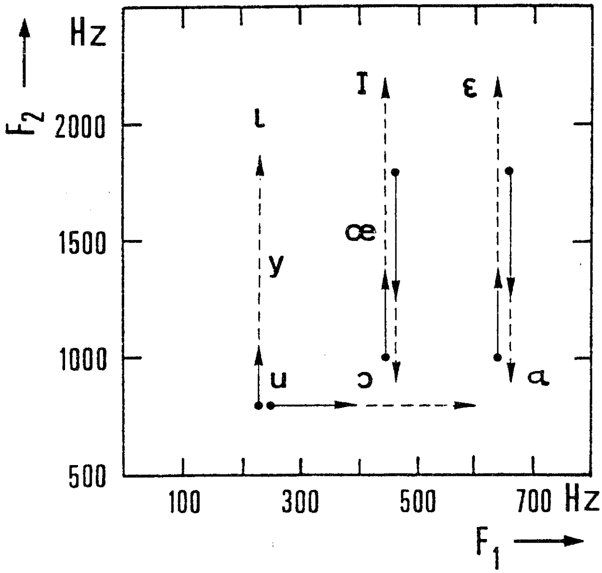

Fig.2. Various formant transitions evaluated so far. Each arrow indicates the smallest transition used in that region, whereas the dashed extension specifies the range over which the transitions were varied. The phonetic symbols roughly indicate the vowel regions.

request. The stationary comparison signal was calculated and generated each time it was requested. When the subject was satisfied about his match of the final part of the test stimulus with the comparison signal, he pressed a button and his setting was stored in the computer, after which the subject could proceed with the next stimulus. Every stimulus was presented twice in a set of 20 stimuli. Performing this task took about three quarters of an hour. On another day the subject evaluated a different set, this time for instance with another F_1 value. Fig.2 presents the transition regions evaluated so far. Several regions had to be excluded since they result in CV- or VC-type stimuli in which the consonantal part sounds like a /w/ or /j/ glide. It is in effect impossible to match such glide-type stimuli with stationary signals.

Although one could argue that for matching itself the vowel quality is irrelevant, we got the impression that vowel categories might nevertheless play a role in the matching results. We therefore also asked our subjects afterwards to label all stationary signals by using orthographic vowel symbols (for Dutch this does not cause much ambiguity). Fig.3a gives these identification results averaged over 12 subjects doing each identification twice.

EXPERIMENTAL RESULTS AND DISCUSSION

For the five dynamic stimuli represented in fig.1 the matching results, averaged over the 12 subjects and the two settings per subject, are given in fig.3b. The deviation from the final frequency is represented as a function of that final or target frequency F_t. The initial frequency of F_2 was always 1800 Hz in this situation. If subjects had been able to match the final frequency, this graph would just show points scattered around zero. However, the deviations seem to follow a regular pattern. Only for the smallest sweep (from 1800 Hz to 1300 Hz) is there an indication of overshoot, so most subjects set the matching stationary vowel to an F_2-value lower than 1300 Hz. For all the other sweeps the matched frequency was

STUDY OF THE ROLE OF FORMANT TRANSITIONS IN VOWEL RECOGNITION USING THE MATCHING PARADIGM

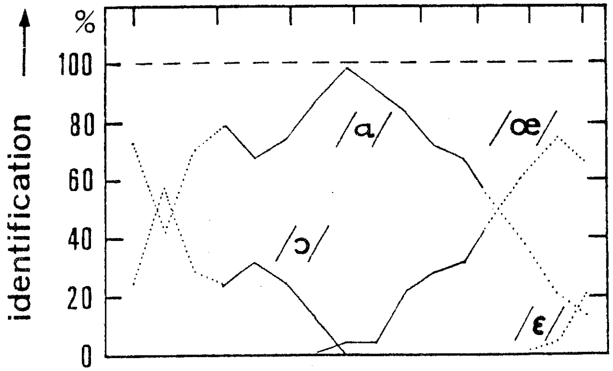

Fig.3a. Identification results for the stationary four-formant vowel signals with F_2-values as indicated.

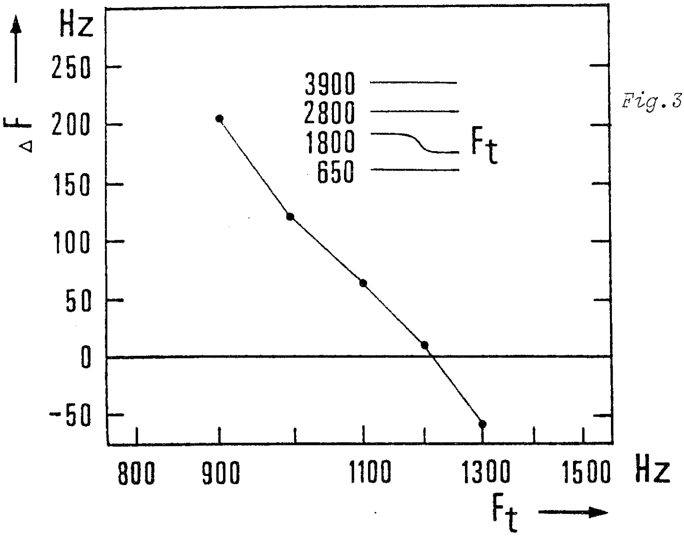

Fig.3b. Matching results for five vowel stimuli for which F_2 drops from 1800 Hz to the indicated target frequency F_t. The difference between matched and target frequency is given along the vertical axis, averaged over 12 subjects and two settings per subject.

higher than the target frequency, which is an indication of undershoot, although one could also interpret this as an averaging mechanism. If this result reflects a purely psychophysical match between a band sweep and a stationary signal, then it would not matter which vowel-like quality this signal represents. We tested this by changing the F_1-value of all stimuli from 650 Hz to 450 Hz leaving everything else the same. This signal sounds like an /oe-I/-type vowel changing towards an /ɔ/-type vowel. The results of this matching experiment are given by the dashed line in fig.4. Apparently these data differ from those of the preceding experiment (continuous line in fig.4). This time the absolute deviation from the final F_2 frequency is smaller, but three of the five sweeps show matching results below the zero line, which indicates overshoot.

Like other researchers [4], we also studied a third condition which could be interpreted as the non-speech analogue. In this experiment the varying stimulus consisted of the same downward sweep of F_2, without however any other formant added,

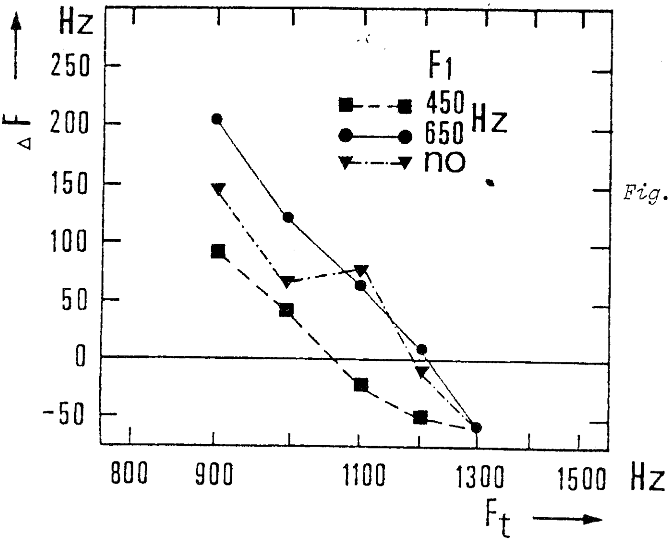

Fig.4. Matching results similar to fig.3b. Two conditions refer to four-formant stimuli with F_1 either 450 or 650 Hz, whereas in the third condition the F_2 transition only is present.

while F_o was also kept constant at 100 Hz. The results of this matching experiment are given in fig.4. by the dash dotted line. Somewhat to our surprise these one-formant stimuli still had a very vowel-like character.
Matching judgments, similar to those in fig.4, were also elicited for dynamic stimuli in which F_2 rises from 1000 Hz to 1400-2200 Hz, again with F_1 equal to 650 Hz, or 450 Hz, or absent. We also did a pilot experiment using an F_1 trajectory rising from 240 Hz to 400-600 Hz, as well as a symmetrical F_1 transition. The data show substantial individual differences although an overall pattern is to be seen. We have not yet been able to evaluate all these data and their implications. It is especially interesting to correlate the matching data with identification data such as those shown in fig.3a.
However, with the experimental matching paradigm that we have made use of up to now, we have not found either the strong and consistent overshoot effects as described by Fujisaki and Sekimoto [4], and Kanamori and Kido [8], or those implicitly present in the identification data of Lindblom and Studdert-Kennedy [11]. More data will be necessary to be able to specify whether these differences are related to an underlying perceptual mechanism, or to the procedure used, or perhaps to specific characteristics of the stimuli.

REFERENCES

[1] P.T. Brady, A.S. House, and K.N. Stevens, "Perception of sounds characterized by a rapidly changing resonant frequency", J. Acoust. Soc. Amer. 33, 1357-1362, (1961).

[2] J.K. Cullen and M.J. Collins, "Audibility of short-duration tone glides as a function of rate of frequency range", Hearing Research 7, 115-125, (1982).

[3] F.M. Danaher, M.J. Osberger, and J.M. Pickett, "Discrimination of formant frequency transitions in synthetic vowels", J. Speech Hearing Res. 16, 439-451, (1973).

STUDY OF THE ROLE OF FORMANT TRANSITIONS IN VOWEL RECOGNITION USING THE MATCHING PARADIGM

[4] H. Fujisaki and S. Sekimoto, "Perception of time-varying resonance frequencies in speech and non-speech stimuli", In: A. Cohen and S.G. Nooteboom (Eds.), Structure and process in speech perception, Springer-Verlag, Berlin, pp. 269-282, (1975).

[5] R.B. Gardner and J.P. Wilson, "Evidence for direction specific channels in the processing of frequency modulation", J. Acoust. Soc. Amer. 66, 704-709, (1979).

[6] J.W. Horst, Discrimination of complex signals in hearing, Ph.D. thesis, University of Groningen, (1982).

[7] A. House and E.P. Neuburg, Unpublished data made available by personal communication in 1983, (1970).

[8] Y. Kanamori and K. Kido, "Perception of vowel stimuli characterized by time-varying formant frequency", J. Acoust. Soc. Japan 32, 277-279, (1976).

[9] F.J. Koopmans-van Beinum, H.A.L. Wouters, H.J.A.G. Buiting, and L.C.W. Pols, "The influence of response categories on the identification of vowels excerpted from conversational speech", this conference, (1984).

[10] H. Kuwabara, "Vowel identification and dichotic fusion of time-varying synthetic speech sounds", Acustica 53, 143-151, (1983).

[11] B.E.F. Lindblom and M. Studdert-Kennedy, "On the role of formant transitions in vowel recognition", J. Acoust. Soc. Amer. 42, 830-843, (1967).

[12] P. Mermelstein, "Difference limens for formant frequencies of steady-state and consonant-bound vowels", J. Acoust. Soc. Amer. 63, 572-580, (1978).

[13] I.V. Nabelek and I.J. Hirsch, "On the discrimination of frequency transitions", J. Acoust. Soc. Amer. 45, 1510-1519, (1969).

[14] M.E.H. Schouten, "Identification and discrimination of sweep tones", presented for publication to Perc. & Psychophysics.

[15] K.N. Stevens and A.S. House, "Perturbations of vowel articulations by consonantal context: An acoustical study", J. Speech Hear. Res. 6, 111-128, (1963).

[16] W. Strange, J.J. Jenkins, and Th.L. Johnson, "Dynamic specification of coarticulated vowels", J. Acoust. Soc. Amer. 74, 695-705, (1983).

[17] B.W. Tansley and J.B. Suffield, "Time course of adaptation and recovery of channels selectively sensitive to frequency and amplitude modulation", J. Acoust. Soc. Amer. 74, 765-775, (1983).