Proceedings of The Institute of Acoustics 4D1

COMBINING FREQUENCY AND DURATION INFORMATION IN VOWEL RECOGNITION.

W. A. AINSWORTH
DEPARTMENT OF COMMUNICATION & NEUROSCIENCE, UNIVERSITY OF KEELE.

It is generally accepted that the frequencies of the first two formants are the most important factors in the recognition of vowel sounds (1, 2, 3). Temporal factors, however, are also important (4, 5, 6, 7, 8, 9). The question therefore arises as to how formant frequency information is combined with duration information in order for a vowel sound to be identified.

The process of the recognition of vowel sounds can be formalised by means of a multidimensional space in which points in that space represent idealised forms of vowels. The dimensions of the space will be the frequency of the first formant (F1), the frequency of the second formant (F2), the duration of the vowel (T), and possibly others such as the bandwidths of the formants and the frequencies of the higher formants. An unknown vowel can also be represented by a point in this space. It is reasonable to suppose that it will be perceived as belonging to that vowel category whose ideal form is represented by the nearest point. The dimensions of the space, however, are not known, nor are the relative effects of differences in formant frequency and differences in duration. The purpose of the present experiments was to investigate these questions.

Method

If the durations of the vowels are deliberately distorted, by making them all very long or very short, errors of identification will be induced which reflect the effects of duration. By studying the error pattern obtained it should be possible to estimate the contribution of duration to vowel recognition.

The stimuli consisted of the eleven synthesised vowels. Their formant frequencies and intensities and their durations were those parameters in a speech synthesis-by-rule programme (10) which were known would be recognised with an accuracy of about 90% by typical British English listeners (11).

They were generated on-line by a synthesis-by-rule system consisting of a parallel-formant speech synthesiser (12) controlled by a minicomputer (13). The listeners were asked to try to identify the sounds they heard as the vowel in one of the words in the set heed, hid, head, had, hard, hod, hoard, hood, who'd, hud, heard; and to press an appropriately labelled switch on a box in front of them. An extra switch was provided in case the sound that they heard did not correspond to any of the vowels in the response set.

In the first session they heard the eleven vowels ten times each in a randomised order in an h-d context preceded by the introductory phrase "the next syllable is", synthesised with fairly natural intonation. In the next session the vowels were presented 20 times in isolation with a steady fundamental of 120Hz and with normal durations. In the next the vowels were presented with a duration of 50 msec, and in the last with a duration of 500 msec. The normal

COMBINING FREQUENCY AND DURATION INFORMATION IN VOWEL RECOGNITION.

range of duration of vowels was 150 msec to 400 msec, so 50 msec was 100 msec less than the shortest, and 500 msec was 100 msec more than the longest.

Results

The 'mean' recognition scores for the 'normal' duration vowels, the 'long' vowels (500 msec), and the 'short' vowels (50 msec) were 87.1%, 79.3%, and 68.4% respectively. The difference between the normal and long vowels was significant ($p < 2.5\%$) by a t-test. The difference between normal and short vowels was significant at $p < 0.05\%$, and the difference between long and short vowels was significant at $p < 1\%$. Hence the durations of the vowels do have a significant effect on their recognition.

Confusion matrices were constructed. With normal vowels the errors were mainly between near neighbours in an F1-F2 space, but with the long vowels there was a tendency for some phonetically short vowels to be heard as phonetically long vowels, and with the short vowels the reverse was true.

Model

The first model to be investigated was one in which an unknown vowel is classified as the nearest ideal vowel in an F1-F2-T space. In order to determine the nearest vowel it is necessary to define a distance measure such as:

$$d_{ij} = (|k_1 \Delta F1|^n + |k_2 \Delta F2|^n + |k_3 \Delta T|^n)^{1/n} \qquad (1)$$

where $\Delta F1 = F1_i - F1_j$, $\Delta F2_i - F2_j$, and $\Delta T = T_i - T_j$.

With n having a value of 1, equation (1) becomes the sum of the differences between two points projected onto the axes. This will be called the linear distance measure (LDM). With n having a value of 2, equation (1) becomes the distance between two points in a three dimensional space, or the euclidean distance measure (EDM).

The appropriate units in which to measure F1 and F2 are not known. The units which were investigated were the physical units (Hz) and the psychophysical units (Barks). A Bark is defined as the width of the critical band at that frequency (14) and may be approximated by

$$f = 650 \sinh (x/7) \qquad (2)$$

where f is the frequency in Hz and x is equivalent psychophysical unit in Barks (15).

In order to test whether an unknown vowel is most likely to be classified as the nearest ideal vowel, it is necessary to define some function which is a measure of this. One such function can be obtained by calculating d_{ij} for each pair of vowels, and then ordering the vowels according to the value of the distance measure. If there are m vowels in the set, define a vector $[O_i]$ such that the nearest point to an unknown vowel has a value m, the next nearest m-1, etc. An order matrix $[O_{ij}]$ can then be constructed for each of the m vowels in the set. A function can then be defined as

$$S = \sum_{ij}^{m} [M_{ij}] * [O_{ij}]/m^2 \qquad (3)$$

where $[M_{ij}]$ is the confusion matrix and $[O_{ij}]$ is the order matrix. Each element

M_{ij} is multiplied by the corresponding element O_{ij} and the results summed. The function S will have a maximum value if the largest elements of each row of M occupy the same positions as the maximum elements of O, or at the minimum distances from the ideal vowels. Any errors of classification will reduce S, and will reduce S by a greater amount the larger the value of the corresponding d_{ij}.

As well as misclassifications induced by distortions of vowel durations, there will also be errors caused by such factors as the listeners' attention wandering, accidentally pressing the wrong switch, or their internal idealised vowels being different from those of the set of test vowels. Some of these factors will result in responses which are dependent on the distance measure, but others will not. In order to allow for such factors another function can be defined

$$S' = \sum_{ij}^{m} [M_{ij}] * [O'_{ij}]/m^2 \qquad (4)$$

where $[O_{ij}']$ is an order matrix derived from the confusion matrix by placing m in the position in each row which contains the largest number, m-1 in the position which contains the next largest, etc. The difference between these two functions

$$\Delta S = S' - S \qquad (5)$$

is then a measure of how well a particular definition of distance measure predicts the confusions which are actually obtained.

Effect of F1 and F2

With normal duration vowels ΔT for the unknown vowel and the perceived vowel should be relatively small, so that equation (1) cam be simplified to

$$dij = (|\Delta F1|^n + |k\Delta F2|^n)^{1/n} \qquad (6)$$

Using this distance measure ΔS was calculated for the LDM and EDM for both the Hz and Bark scales of frequency for k ranging from 0 to 2. The results showed that ΔS is minimised for $k \geq 1$, and that neither the distance measure nor the frequency scale had any noticeable effect on the result.

Effect of duration

As F1 andF2 appear to have approximately equal effects, equation (1) can be simplified to:

$$d_{ij} = (|\Delta F1|^n + |\Delta F2|^n + |k\Delta T|^n)^{1/n} \qquad (7)$$

The next task is to estimate k in order to discover the relative effects of formant frequency and vowel duration.

The function ΔS was calculated for the normal duration vowels, the short vowels (50 msec) and the long vowels (500 msec) for a range of values of k. These were added together and plotted against k for the LDM and the EDM for both the Hz and Bark scales. The results showed that for the LDM a minimum occurs at $k = 0.4$ for the Hz scale and 0.7 for the Bark scale. With the EDM, duration has no effect for the Hz scale over a range of k of 0 to 0.4, and for the Bark scale the minimum occurs at $k = 0.4$. The lowest minimum occurs for the EDM with the Hz scale, but the differences between the various measures are small.

COMBINING FREQUENCY AND DURATION INFORMATION IN VOWEL RECOGNITION

References

1. P. DELATTRE, A.M. LIBERMAN, F.S. COOPER and L.J. GERSTMAN 1952 Word 8, 195-210. An experimental study of the acoustic determinants of vowel color.
2. G.E. PETERSON and H.L. BARNEY 1952 J. Acoust. Soc. Am. 67, 276-287. Control methods used in a study of the vowels.
3. L.W.C. POLS, L.J. VAN DER KAMP and R. PLOMP 1969 J. Acoust. Soc. Am. 46, 458-467. Perceptual and physical space of vowel sounds.
4. K.N. STEVENS 1959 J. Acoust. Soc. Am. 31, 109(a). Effect of duration on identification.
5. A. COHEN, I.H. SLIS and J. T'HART 1967 Phonetica 16, 65-70. On tolerance and intolerance in vowel perception.
6. D.C. BENNETT 1968 Language and Speech 11, 65-85. Spatial form and duration cues in the recognition of English and German vowels.
7. W.A. AINSWORTH 1972 J. Acoust. Soc. Am. 51, 648-651. Duration as a cue in the recognition of synthetic vowels.
8. D.H. KLATT 1976 J. Acoust. Soc. Am. 59, 1208-1221. Linguistic uses of segmental duration in English: acoustic and perceptual evidence.
9. S.G. NOOTEBOOM and G.J.N. DOODEMAN 1980 J. Acoust. Soc. Am. 67, 276-287. Production and perception of vowel length in spoken sentences.
10. W.A. AINSWORTH 1974 Int. J. Man-Mach. Stud. 6, 493-511. Performance of a speech synthesis system.
11. W.A. AINSWORTH 1978 J. Acoust. Soc. Am. 63, 1528-1534. Perception of speech sounds with alternate formants presented to opposite ears.
12. J.N. HOLMES, I.G. MATTINGLY and J.N. SHEARME 1964 Language and Speech 7, 127-143. Speech synthesis by rule.
13. W.A. AINSWORTH and J.B. MILLAR 1981 Behav. Res. Methods Instrum. 3, 21-24. A simple time-sharing system for speech perception experiments.
14. E. ZWICKER and H. FASTL 1972 J. Acoust. Soc. Am. 52, 699-702. On the development of the critical band.
15. M.R. SCHROEDER, B.S. ATAL and J.L. HALL 1979 Frontiers of Speech Communication Research, B. Lindblom and S. Öhman (Eds), 217-229, Academic Press, London. Objective measure of speech signal degradations based on masking properties of human auditory perception.