SPEEDED DETECTION OF VOWELS AND STEADY-STATE CONSONANTS

Dennis Norris[1], Brit van Ooyen[2] and Anne Cutler[1]

[1]MRC Applied Psychology Unit,
15 Chaucer Rd., Cambridge, CB2 2EF, United Kingdom.

[2]Dept. of Linguistics/Phonetics Laboratory,
University of Leiden, P.O. Box 9515, The Netherlands.

ABSTRACT

We report two experiments in which vowels and steady-state consonants served as targets in a speeded detection task. In the first experiment, two vowels were compared with one voiced and once unvoiced fricative. Response times (RTs) to the vowels were longer than to the fricatives. The error rate was higher for the consonants. Consonants in word-final position produced the shortest RTs. For the vowels, RT correlated negatively with target duration. In the second experiment, the same two vowel targets were compared with two nasals. This time there was no significant difference in RTs, but the error rate was still significantly higher for the consonants. Error rate and length correlated negatively for the vowels only. We conclude that RT differences between phonemes are independent of vocalic or consonantal status. Instead, we argue that the process of phoneme detection reflects more finely grained differences in acoustic/articulatory structure within the phonemic repertoire.

1. INTRODUCTION

When listeners misperceive speech, vowels are less affected than consonants. In particular, vowels in stressed syllables are accurately perceived [1]. Assuming that this apparent resilience to perceptual distortion can be explained in terms of their relative prominence in the acoustic signal, one could conclude that vowels are easier to perceive than consonants. However, in the phoneme detection task, in which listeners' response time to detect the presence of a pre-specified phoneme target is measured, vowels produce longer response times than consonants, suggesting that they are harder to perceive. Cutler, Norris and van Ooyen |2| carried out two detection experiments with English real words and nonwords. The results showed long RTs for vowels compared to RTs as obtained for consonants in previous work. Full vowels were responded to faster and more accurately than reduced vowels. Vowel duration correlated negatively with RT. In a subsequent experiment, van Ooyen, Cutler and Norris [3] compared RTs to the vowels /a/ & /i/ on the one hand and to the consonants /p/ & /t/ on the other in the same subject population. Targets occurred in word initial, medial and final position. Again, the vowel targets yielded longer RTs than the consonant targets in all word positions, although the difference was minimal in word initial position. RTs to the vowels correlated negatively with target length: longer vowels were responded to faster. Taken together, the above results were seen as evidence that, in English, speeded detection of vowels is harder than that of consonants.

Strictly speaking, the distinction between vowels and consonants is based on phonological factors. Vowels form syllabic nuclei whereas consonants occur optionally and in varying numbers at syllable margins. In a follow-up study, van Ooyen, Cutler and Norris [4] concentrated on this difference in syllabic function as a possible explanation for the observed RT differences. To enable as pure a test as possible of syllabic function alone, the vowels /i/ & /u/ were compared to the consonants which most closely resembled them in acoustic/articulatory characteristics, namely the semivowels /j/ & /w/, respectively. This time, the vowels were responded to both faster and more accurately than the semivowels, just as in the previous experiment [3] vowels had produced longer RTs than stops. It was concluded that differences in detection performance are not related to differences in syllabic function between vowels and consonants. However, there are acoustic/articulatory correlates of the phonological distinction. In general, vowels are characterised as relatively long, steady-state, periodic sounds, and consonants as relatively transient, aperiodic sounds. Seeing that the observed RT differences bear no relation to the phonological distinction, closer investigation of these acoustic/articulatory correlates seems warranted.

In sum, unvoiced stops are detected fastest, followed by vowels, followed by semivowels. Furthermore, vowels show a systematic negative correlation of RT with target phoneme length, stops and semivowels do not. However, we cannot conclude from this that such a correlation is unique to vowels and therefore indicative of a processing difference between vowels and consonants. Unlike vowels, both stops and semivowels take up comparatively short portions of the acoustic signal. In our previous experiments [3,4], target lengths for the vowels varied between 108-382 ms, for the stops and the semivowels combined between

11-109 ms; there is virtually no overlap between the two sets of measurements. It is conceivable that the range of duration for the consonants is simply too small to show significant correlations of length with RT. Extrapolating from the observed negative correlation for vowels though, if anything we would predict long RTs for these short consonants. We therefore decided to carry out a direct comparison of the relationship between detection RT and target duration. The same two vowels /a/ & /i/ from our previous experiment [3] were used in combination with consonants which still share some of the acoustic/articulatory characteristics of the consonants used so far, but which at the same time resemble vowels in terms of their comparatively long, maintainable (as opposed to momentary) character. We call these consonants 'steady-state'. Experiment 1 compares the vowels /a/ & /i/ with two fricatives /s/ and /v/, which resemble stops in that they are classified as obstruents 15]. That is, they are produced with a cavity configuration that makes spontaneous voicing impossible. On the other hand, just like vowels, fricatives are classified as continuant. Experiment 2 compares the same vowels /a/ & /i/ with two nasal consonants /m/ & /n/, which share with stops the fact that their primary constriction results in a complete blocking of the airflow from the lungs. On the other hand, just like vowels (and semivowels) these are sonorant sounds, that is they are produced with spontaneous voicing. If acoustic/articulatory characteristics of phonemes can provide cues to their ease of recognition in a speeded detection task, we expect that both the fricatives and the nasals will produce RTs in between those previously obtained for momentary consonants and simple vowels, on account of the fact that they share consonantal as well as vocalic characteristics. Specifically, if target duration is involved in this task, we expect the steady-state consonants to produce an RT pattern similar to that of vowels.

2. EXPERIMENT 1

Method

Materials. The target vowels were high front /i/ and low back /a/ plus the voiced fricative /v/ and the unvoiced fricative /s/. The materials for the vowels were the same physical stimuli as those used in a previous experiment [3]. The four target phonemes occurred in 144 mono- and disyllabic words, 36 for each target phoneme. Of these 36 words, 12 had the target phoneme in word initial, 12 in medial and 12 in final position. For /a/ & /i/, 20 of the 36 words were monosyllabic, 16 were disyllabic. For /s/ & /v/, 12 of the 36 words were monosyllabic, 24 were disyllabic. Half of all disyllabic words had initial stress, half had final stress. Target phonemes always occurred in stressed syllables. The words were matched for frequency for initial, medial and final means within each target phoneme. Between target phonemes, /a/ was matched with */s/* and /i/ with /v/. Forty mono- and disyllabic words, 10 per target phoneme, were dummy target items. About 1000 mono- and disyllabic words were filler items.

Experimental design. The material formed 4 blocks, one per target phoneme. Each block consisted of 55 lists of 2 to 6 words in length. Of these, 36 lists contained an experimental item in 3rd, 4th or 5th position, 10 lists had a dummy target item in 1st or 2nd position and 10 lists contained no occurrence of the target phoneme. Before each block the target phoneme was specified with examples. The blocks, together with the examples, a short practice block and instructions, were recorded by a male native speaker of British English. The blocks were presented in four different orders.

Subjects. Twenty-four subjects between 18 and 25 years of age were paid participants. All were native speakers of British English with normal hearing. Six subjects heard each order of presentation of the blocks.

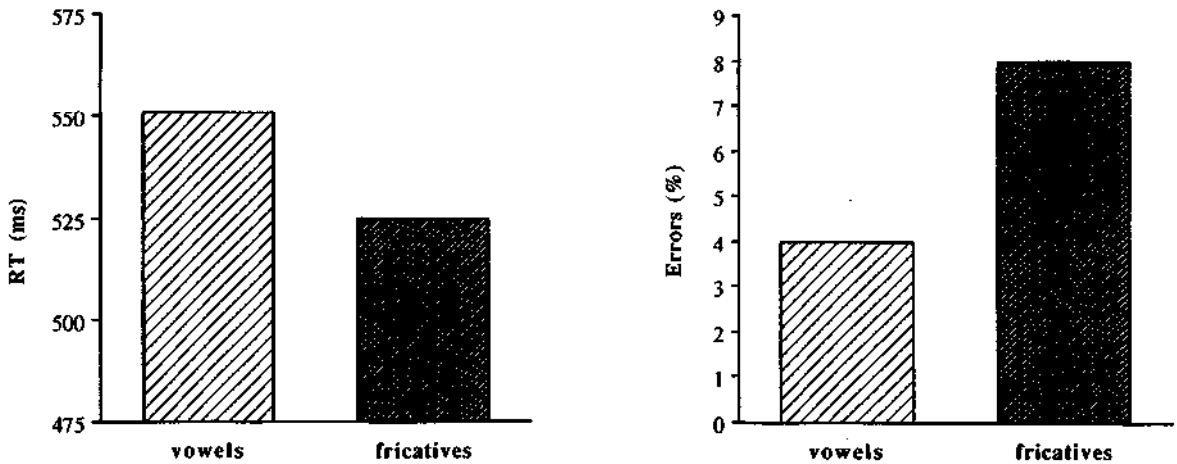

Figure 1. Mean RTs (ms) and error rates (%) for vowels and fricatives. Experiment 1.

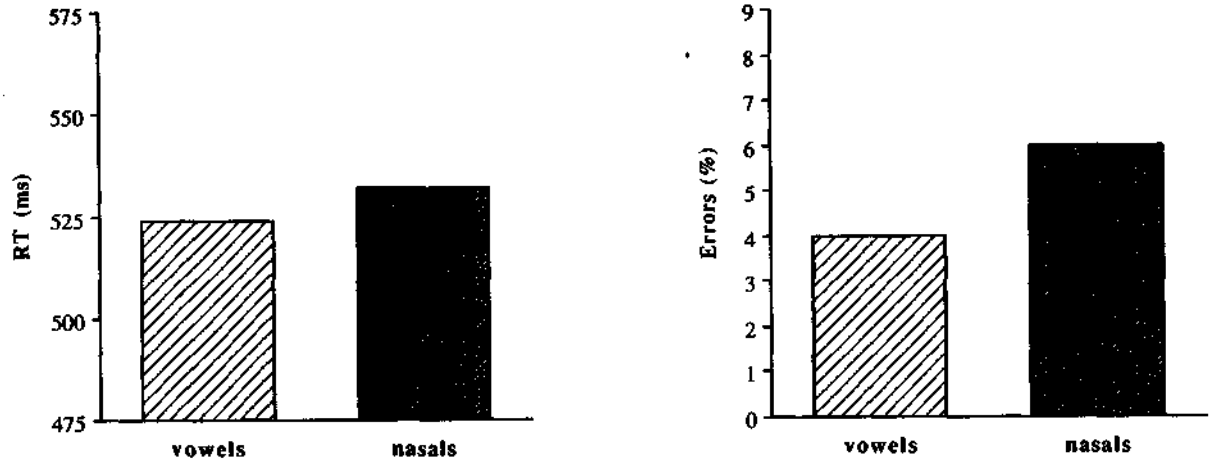

Figure 2. Mean RTs (ms) and error rates (%) for vowels and nasals. Experiment 2.

Procedure. Subjects were tested individually. They listened to taped instructions that requested them to press a single response key as soon as they detected a target phoneme, as specified in the examples, anywhere in a word. A timing mark, aligned with the onset of each experimental word, initiated response timing. The data were stored on a microcomputer. The 144 experimental words were digitized to measure word length, target phoneme duration, and time between target phoneme onset and timing mark. RTs were adjusted for these values to give RTs from phoneme onset.

Results

Response times below 100 ms or above 1500 ms were discarded. Analyses of variance with subjects and with words as random factors showed a marginally significant RT advantage for the consonants (Fl [1,20] = 3.75, $p < .07$; F2 [1,132] = 13.35, $p < .01$). Fig. 1 shows mean RTs in ms and error rates in % for the vowels (551 ms, 4%) versus the fricatives (525 ms, 8%). There was a significant main effect of position of the target in the word (Fl [2,40| = 35.63, $p < .001$; F2 [2,132] = 50.15, $p < .001$); but this effect interacted with the vowel-consonant factor (Fl [2,40] = 51.60, $p < .001$; F2 [2,132] = 32.16, $p < .001$). T-tests showed that mean RTs to word-medial vowels (581 ms) were significantly longer than those to either word-initial (538 ms; tl [23] = 2.27, $p < .04$; t2 [47] = 3.36, $p < .01$) or word-final vowels (535 ms; tl [23] = 4.63, $p < .01$; t2 [47] = 3.80, $p < .01$). There was quite a different pattern for the consonants whereby word-medial targets (549 ms) yielded significantly shorter mean RTs than word-initial targets (601 ms; tl [23] = 3.66, $p < .01$; t2 [47] = 4.59, $p < .01$) and word-final targets (424 ms) had significantly shorter mean RTs than word-medial (tl [23] = 8.24, $p < .01$; t2 [47| = 10.32, $p < .01$) targets. This suggests a lexicality effect for the consonants only.

The overall error rate was 6%, with consonants at 8% being missed significantly more often than vowels at 4% (Fl [1,20] = 12.44, $p < .01$; F2 [1,132] = 18.37, $p < .001$). A negative correlation was found between RT and measured duration for the vowels (r [71] = -.29, $p < .02$), but not for the consonants. Error rate correlated negatively with duration both for vowels (r [71] = -.41, $p < .01$) and for fricatives (r [71] = -.28, $p < .02$).

3. EXPERIMENT 2

Method

Materials and design. Target phonemes were again high front /i/ and low back /a/ plus the nasals /m/ and /n/. This time, /n/ was matched in frequency with /a/, and hence also with /s/ from Experiment 1, and /m/ with /i/, and hence also with /v/ from Experiment 1. The experimental design was otherwise as for Experiment 1.

Subjects. Twenty-four native speakers of British English between 19 and 26 years of age were paid participants. All reported normal hearing. Six heard each order of presentation of the blocks.

Procedure. This was as for Experiment 1.

Results

This time, the mean RT for the two vowels at 524 ms did not differ significantly from the mean RT for the two consonants at 532 ms. Fig. 2 shows mean RTs in ms and error rates in % for the vowels versus the nasals. Again, there was a significant effect of position of the target in the word (Fl [2,40] = 42.63, $p < .01$; F2 [2,132] = 42.56, $p < .01$), which once more interacted with the vowel-consonant factor (Fl [2,40] = 28.92, $p < .01$; F2 [2,132] = 18.05, $p < .01$). T-tests showed that mean RTs to word-medial vowels (548 ms) were significantly longer than those to word-final vowels (505 ms; tl [23] = 3.35, $p < .01$; t2 [47] = 3.85, $p < .01$) and also significantly longer than those to word-initial vowels (519 ms; tl [23] = 2.49, $p < .02$; t2 [47] = 3.36, $p < .01$). Once more, a different pattern emerged for the consonants such that word-final targets (439 ms) had significantly shorter mean RTs than either word-initial (573 ms; tl [23] = 9.69, $p < .01$; t2 [47] = 10.92,

$p < .01$) or word-medial (583 ms; tl [23] = 8.70, $p < .01$; t2 [47] = 12.51, $p < .01$) targets.

The overall error rate was 5%, with consonants at 6% being missed significantly more often than vowels at 4% (Fl [1,20] = 5.11, $p < .05$; F2 [1,132] = 4.16, $p < .05$). Correlations between RT and measured duration were not significant. A correlation analysis of error rate with duration however showed again a negative correlation for the vowels (r [71] = -.42, $p < .01$), but this time not for the consonants.

4. CONCLUSION

The two experiments have shown that English vowels are somewhat harder to detect than fricatives, but not harder than nasals. In our previous experiment [3] which used the same set of vowel targets, these were much harder to detect than stop consonants, but in another previous experiment [4] they were easy relative to semivowels. So some consonants produce similar RTs to vowels, suggesting that vowels do not stand out on their own as taking more time to recognise. There are, however, several other ways in which they do seem to stand out. First, vowels show systematically lower error rates than consonants, so they are perceived more accurately. Second, only vowels have systematic negative correlations beween RT and target duration, so longer ones are easier to recognise. Also, vowels show negative correlations between error rate and duration, so longer ones are perceived more accurately. For nasals there is no such correlation. For fricatives there is, but it is not tied in with speed of response. In other words, only detection of vowels can be speeded up by more information in their steady-state portion.

Independent evidence that durational information for vowels is important comes from both human and automatic speech recognition studies. Ainsworth [6] presented synthesized vowels with a wide range of duration for identification and found that duration was an important cue to vowel identity. In experiments on the detectability of isolated synthetic vowels, Kewley-Port [7] showed that vowel thresholds decreased with increased vowel duration. Lastly, Deng et al. [8] report on a hidden Markov model-based recognition system the performance of which was improved with the use of vowel durational models.

It would seem that, even though intrinsic vowel duration is not used contrastively in English, there is a relationship between physical duration and the process of sufficiently identifying a stimulus input in order for a matching response to be made. Recalling the observation that vowels form a continuum, this implies that there is more scope for vowel targets to be ambiguous than there is for consonants. In theory it is possible to produce a vowel at any specified distance between any two other vowels in the spectrum. This makes it much harder to perceive any given vowel stimulus as an unambiguous token of a particular type. This in turn suggests that a listener will benefit from having as much information as possible in vowel detection. This idea is consistent with the finding that in both Experiment 1 and 2, only the vowels show a systematic disadvantage for targets in word-medial position. Here the vowel is most open to coarticulatory influences, rendering it more ambiguous.

Three things are worth noting. First, the studies just discussed all used synthetic stimuli. Second, no direct comparison was made between relative importance of durational information for vowels versus consonants. Third, unlike these studies, our methodology is always the speeded detection task. It may well be that, for example, identification tasks pose quite different requirements on the processing mechanism from the on-line, speeded detection task. Similarly, it may well be that analyses of spontaneous slips of the ear [1] tap into a different level of speech processing than the speeded detection task. Nevertheless, the finding that vowels are perceived more accurately than consonants is robust across all of these studies. Thus, these results can serve as an extension of previous findings to an on-line, direct comparison of vowel and consonant perception.

5. ACKNOWLEDGEMENTS

This research was supported by the ESPRIT BRA program [project P3207].

6. REFERENCES

[1] Z.S. Bond & S. Games, "Misperceptions of Fluent Speech", in R. Cole (Ed.) *Perception and Production of Fluent Speech,* Hillsdale, NJ: Erlbaum, 1980.

[2] A. Cutler, D. Norris & B. van Ooyen, "Vowels as Phoneme Detection Targets", *Proceedings of the International Conference on Spoken Language Processing,* Kobe, Japan, Vol. 1, pp. 581-584, 1990.

[3] B. van Ooyen, A. Cutler and D. Norris, "Detection Times for Vowels versus Consonants" *Proceedings of Eurospeech '91,* Genova, Italy, Vol. 3, pp. 1451-1454, 1991.

[4] B. van Ooyen, A. Cutler and D. Norris, "Detection of Vowels and Consonants with Minimal Acoustic Variation", *Speech Communication,* Vol 11, 1992, in press.

[5] N. Chomsky and M. Halle, *The Sound Pattern of English,* New York: Harper & Row, 1968.

[6] W. Ainsworth, "Duration as a Cue in the Recognition of Synthetic Vowels", *Journal of the Acoustical Society of America,* Vol. 51, pp. 648-651, 1972.

[7] D. Kewley-Port, "Detection Thresholds for Isolated Vowels", *Journal of the Acoustical Society of America,* Vol. 89, pp. 820-829, 1991.

[8] L. Deng, M. Lennig and P. Mermelstein, "Use of Vowel Duration Information in a Large Vocabulary Word Recognizer", *Journal of the Acoustical Society of America,* Vol. 86, pp. 540-548, 1989.