On Accounting for the Poor Recognition of Isolated Vowels*

Donald Shankweiler, + Winifred Strange, ++ and Robert Verbrugge +++

ABSTRACT

Earlier studies have shown that vowels spoken in isolation tend to be poorly perceived, even when they are produced by phonetically trained talkers. Listeners, however, generally make remarkably few errors in identification of vowels in consonant-vowel-consonant (CVC) environment, even when each syllable is uttered by a different talker. Sets of nine American English vowels were spoken by a panel of talkers: in isolation and in a /p-p/ environment. Measurements of the first three formant frequencies were obtained from spectrograms. Listening tests were made up by randomizing talkers and tokens and these were presented to phonetically naive listeners. Percent recognition of the intended vowel (averaged over vowels) was 83 percent for the /p-p/ condition and 58 percent for the isolated condition. When the asymptotic formant frequencies of each talker’s isolated and medial vowels are compared, the values are found to be highly similar. A nontrivial explanation must be sought for the perceptual difficulty of isolated steady-state vowels. The data point to the conclusion that no single temporal cross section of a syllable conveys as much vowel information to a perceiver as is given in the dynamic contour of the formants.

Central to current conceptions of the vowel is the idea of the target. In articulatory terms the target is a configuration of the vocal tract toward which the articulators aim. In practice, ideal vowel targets are defined in the acoustic record by formant frequency values obtained from quasi steady-state vowels produced in isolation. It is well-known, of course, that in words and sentences steady states are rarely attained and that formant frequencies usually

*Paper presented at the 89th meeting of the Acoustical Society of America, Austin, Texas, 10 April 1975.

+ Also University of Connecticut, Storrs.

++ University of Minnesota, Minneapolis.

+++ University of Michigan, Ann Arbor.

Acknowledgment: The research reported here is a cooperative endeavor shared by the Center for Research in Human Learning of the University of Minnesota, and Haskins Laboratories. It is supported in part by grants to the Center and to Haskins Laboratories by the National Institute of Child Health and Human Development.

[HASKINS LABORATORIES: Status Report on Speech Research SR-42/43 (1975)]
vary as a function of time throughout the vowel. Consonantal environment produces systematic shifts in formant frequencies causing them to deviate from target values in direction and amount that is largely predictable from articulatory considerations (Lindblom, 1963; Stevens and House, 1963). It is not known how perceivers take account of context-conditioned variation in perception of vowels. Generally it is assumed that the listener extracts the target formant values whether or not these are acoustically realized in the signal. This view encounters a difficulty, however. It has been noted several times in the literature that isolated vowels tend to be poorly perceived. This raises the question of whether vowels can be adequately described by a compact table containing only a single value for each of the first two or three formants.

In a previous study (Strange, Verbrugge, and Shankweiler, 1974), we investigated the contribution of consonantal environment to perception of medial vowels. Nine American English vowels were produced in a number of consonantal environments and in isolation by a panel of 15 untrained talkers, which included five children—ranging in age from four to ten—five adult females, and five adult males. The utterances were recorded on magnetic tape and assembled into a set of listening tests by randomly mixing the voices from token to token. We presented the tests to a group of listeners for whom these were novel voices. Misidentification of isolated vowels occurred with significantly greater frequency than of medial vowels in a number of consonantal environments. Here we present the results of perception tests for vowels in /p-p/ environment and in isolation. An average of 17 percent of the vowel nuclei were misidentified as the talker's intended vowel when they occurred in the /p-p/ frame, and 42 percent were misidentified in isolation. Figure 1 shows a vowel-by-vowel breakdown of the errors.

It is apparent that the presence of a consonantal environment produced a consistent facilitation in identification of all nine vowels. Our data are in agreement with earlier findings of Fairbanks and Grubb (1961), Fujimura and Ochiai (1963), and Lehiste and Meltzer (1973). We can conclude that isolated vowels are significantly more often misidentified than medial vowels spoken under comparable conditions. Could it be that the acoustic complexities introduced by syllabic structure better serve the requirements of the perceptual apparatus than do quasi steady-state targets?

Before accepting this conclusion, we must first ascertain whether the talkers produced isolated vowels with formant frequencies uncharacteristic of the values reported by earlier investigators. Since isolated vowels do not typically occur in natural speech, we cannot overlook the possibility that our talkers, who had no training in phonetics, produced them in peculiar ways that rendered them relatively unintelligible to the listeners.

The purpose of this study was to investigate that possibility. We undertook spectrographic analysis of the tokens of isolated vowels and medial vowels used in our listening tests. Spectrograms were made on a voiceprint spectrum analyzer of the tokens used in the perceptual tests. Sections were made at the point of closest approach to steady state. Center frequencies of F₁, F₂, and F₃ were measured. Tokens uttered by children and women were first rerecorded at half-speed to facilitate the task of locating the formants. We turn first to the results for isolated vowels.
Figure 1: Mean percent errors in identification of each of nine vowels in isolation and in /p-p/ environment for 15 talkers, randomly mixed (from Strange et al., 1974).
In Figure 2, we see $F_1/F_2$ plots of each of five tokens of the nine vowels for the five children. The enclosed areas include all the tokens of a given type. Asterisks give the average values based on child talkers from the Peterson and Barney (1952) survey. The Peterson–Barney values make an appropriate standard of comparison inasmuch as Stevens and House (1963) have demonstrated that asymptotic formant frequencies of vowels in /h-d/ environment, as employed by Peterson and Barney, closely approximate those obtained for vowels in isolation. The figure shows that, with the exception of /æ/ and /ɔ/, our measurements cluster around the Peterson–Barney average values.

Figures 3 and 4 give the data for adult females and males, respectively. For all three groups of talkers, when the tokens are segregated by category of speaker, the vowels separate out well, except for the women's back vowels. The values for /ɔ/ tend to be displaced in the direction of /a/, reflecting the predominant dialect of the upper Midwest, which minimizes the /a/-/ɔ/ distinction. But, in general, it is apparent that the target values attained by our talkers in production of isolated vowels agree rather well with the values reported by Peterson and Barney (1952) for vowels in an /h-d/ environment. Thus, we can conclude that our talkers, for the most part, adopted conventional targets in their productions of isolated vowels.

Figure 5 depicts the vowel spaces for children, women, and men based on average values of $F_1$ and $F_2$ for each vowel. These vowel polygons show characteristic differences for the three categories of talkers.

The next step was to compare directly the talkers' vowel spaces for isolated vowels and for medial vowels in /p-p/ environment. Figure 6 shows that the spaces are largely congruent. $F_2$ of the medial vowels showed a slight migration toward the center of the space. This is in accord with results reported by Stevens and House (1963) for vowels produced between labial consonants. The effect of the shift is to reduce the acoustic contrast among the medial vowels relative to the isolated vowels.

Figure 7 displays all the tokens actually used on the listening tests arrayed in $F_1/F_2$ space. Isolated vowels are displayed in black; medial vowels in gray. Here it is abundantly clear that the sets of vowel tokens produced under the two conditions of this experiment occupy approximately the same space. The isolated vowels are, as expected, a little better separated acoustically than the vowels in /p-p/ environment.

It should be mentioned that for both sets of vowels a proportion of the tokens deviates markedly from the average values. Before we accept the conclusion that the inferior intelligibility of isolated vowels cannot be attributed to aberrant formant values, we should ask whether error rates on individual tokens can be predicted from their acoustic distance from an average value. If this is the case, then it becomes important to establish whether the variability of targets attained for a given vowel is greater for isolated vowels than for vowels in /p-p/ environment. A full answer to these questions awaits further study. We know, however, that the relative intelligibility of a token cannot be estimated very precisely from its position in the space defined by the two formants, a fact also noted by Peterson and Barney (1952).

Likewise, measurements of vowel duration indicate that differences in durations of medial vowels and isolated vowels fail to account for the consistent
Figure 2: F₁/F₂ plots of five tokens of nine English vowels spoken by five children.
Figure 3: $F_1/F_2$ plots of five tokens of nine English vowels spoken by five adult females.
ISOLATED VOWELS: FIVE MEN

Figure 4: $F_1/F_2$ plots of five tokens of nine English vowels spoken by five adult males.
Figure 5: Mean $F_1/F_2$ points for five tokens of nine English vowels averaged separately for men, women, and children.
Figure 6: Mean $F_1/F_2$ points for five tokens of nine English vowels in /p-p/ environment and in null (θ-#) environment spoken by 15 talkers, randomly mixed.
MIXED TALKER TESTS

ISOLATED VOWELS
MEDIAL P-P VOWELS

Figure 7: $F_1/F_2$ plot of 45 vowel tokens spoken in /p-p/ environment and in null (/#/) environment by 15 talkers.
increase in error rate for all isolated vowels. Although the durations of the latter were considerably longer than vowels in /p-p/ environment, the relative durations of the isolated vowels were the same as for vowels in /p-p/ position, except for /u/, /e/, and /i/, which were relatively longer in isolation than in context.

Having tentatively ruled out a theoretically uninteresting explanation of the difference in intelligibility between these two sets of vowels, we may ask what accounts, then, for the superior intelligibility of vowels in the stop environment. Surely this is puzzling, given the usual assumptions about the nature of the acoustic information that specifies vowel quality. An isolated, quasi steady-state utterance in which the formants attain appropriate targets ought to be an optimal signal for perception. Indeed, synthetic steady-state vowels based on these formant parameters are fairly intelligible to listeners. Moreover, in the domain of automatic speech recognition, some success has been achieved with a static model of the vowel. Gerstman (1968) devised an algorithm based on values of $F_1$ and $F_2$ derived from spectrographic measurements of center formant frequencies of /h-d/ syllables recorded from 76 talkers by Peterson and Barney (1952). Gerstman's algorithm sorted nine vowels in this set with only 2.5 percent error, less than was made by human listeners. From such a result, one might infer that target formant frequencies can in principle unambiguously specify the vowels of English as produced by a variety of talkers. It is but a short step to the conclusion that a human listener's strategy in identifying vowels is to extract the target formant values by means of something like a filter bank.

However, as we saw, this conception of the vowel cannot be reconciled easily with certain facts of perception. Since vowels in isolation were poor signals from the perceiver's standpoint, even though talkers adopted appropriate targets (differing little from /p-p/ targets), we can conclude that target frequencies do not adequately specify a vowel. Cues that we ordinarily regard as consonantal must contribute to the perception of the vowel. We suspect that much vowel information is contained in formant transitions, as Lindblom and Studdert-Kennedy (1967) suggested some time ago. In an analysis of perceptual adjustments for differences in stress and speaking rate, these investigators found that vowel identifications varied with direction and rate of transitions even when the formant frequency values at syllable centers were held constant. A case for the importance of formant transitions in vowel perception might also be made from considerations, such as those raised by Liberman (1970), of the extent of variation in formant contours conditioned by phonetic context. In any case, we are planning experiments to test the hypothesis directly by studying the effects of different consonantal environments, with and without transitions, on the perception of a coarticulated vowel.

Whatever the nature of the contribution consonantal environment makes to the identification of a vowel, the data point to the general conclusion that no single temporal cross section of a syllable conveys as much vowel information to a perceiver as is given in the dynamic contour of the formants. Thus it would seem that the definition of a vowel, from the standpoint of perception, ought to include a specification of how the relevant acoustic parameters change over time. If this conclusion is correct, then the specification of the relation between sound and percep presents the same problems for vowels as for consonants.
REFERENCES


