What Information Enables a Listener to Map a Talker's Vowel Space?

Robert Verbrugge, Winifred Strange, and Donald Shankweiler

The acoustic structure of vowels varies markedly from one speaker to another, and from one phonemic context to another (Peterson and Barney, 1952; Peterson, 1961; Stevens and House, 1963). It is commonly assumed that a listener identifies a talker's vowels in terms of the relation between their acoustic structure and the acoustic structure of other vowels produced by the same person (Joos, 1948; Ladefoged and Broadbent, 1957; Ladefoged, 1967). It is also common to speak of vowels as situated in a vowel space, the shape of which is a function of each individual's vocal tract characteristics. As a result, a talker's vowel space would be completely specified only after a listener hears an extended sample of the talker's speech. Such experience would allow the listener to calibrate or normalize to each particular voice he encounters. This suggests that the largest source of errors in identifying vowels will be inadequate exposure to a novel voice, i.e., hearing an utterance that is too brief or impoverished to allow accurate calibration.

Our first study was an attempt to assess this claim quantitatively, by comparing the identification of vowels under two conditions. In the Mixed Condition a large number of talkers spoke a series of syllables; on any one syllable the listener encountered a voice that was unfamiliar and unpredictable. In the Blocked Condition subjects heard the same series of syllables spoken by one person, so there was ample opportunity to become familiar with the voice, and the talker was fully predictable from one syllable to the next.


† University of Minnesota, Minneapolis.

‡‡ Haskins Laboratories, New Haven, Conn., and University of Connecticut, Storrs.

Acknowledgment: This paper and the following one report a portion of research begun during the academic year 1972-73 while D. Shankweiler was a guest investigator at the Center for Research in Human Learning, University of Minnesota, Minneapolis. The work was supported in part by a grant from the National Institute of Child Health and Human Development to the Center, and in part by grants awarded to Shankweiler and to J. J. Jenkins by the National Institute of Mental Health. We wish to thank Thomas Edman and Kevin Jones for their assistance in every phase of the experimental work and James Jenkins for his advice and encouragement.

Nine vowels appeared in a fixed consonant frame, /p-p/, to form the following syllables: /pIp, pTp, pEp, pAp, pOp, pAp, pOp, pUp/. Each of the nine syllables was spoken five times, for a total of 45 tokens per test.

In the Mixed Condition, 15 talkers were chosen: 5 men, 5 women, and 5 children, representing a wide variety of vocal tract sizes and fundamental frequencies. Each of the 15 people spoke three different vowels during the test. The three tokens for each talker were separated by at least eight other talkers. Listeners heard the test twice, making a total of 90 judgments, 10 for each vowel. They recorded their judgments by circling the appropriate word on an answer sheet.

In the Blocked Condition, a representative man, woman, and child each spoke the full series of 45 test items. Listeners heard each of the three tapes, in one of several orders. Data for only the first two repetitions were pooled together across groups to keep the scores comparable to the Mixed Condition, i.e., five judgments per vowel in a first repetition and five judgments in a second repetition. A judgment was considered an "error" if the indicated vowel was placed in a different phonemic category than that intended by the experimenters. The error measure, then, is a compound of talker and listener processing. In this account of our experiments, we do not attempt to separate these sources.

Listeners made an average of 17 percent errors in identifying vowels produced by the panel of randomly ordered talkers (the Mixed Condition), while in the Blocked Condition, listeners averaged 9.3 percent errors for the vowels of the three single talkers. Thus, it is plain that familiarity with a talker's voice significantly improved the accuracy of identification, though less than half of the errors can be attributed to this source.

There are two ways to look at these error percentages. First, 9 percent is a relatively high "error" rate, considering the complete predictability from trial to trial of both the speaker's voice and the consonantal frame; there are sources of vowel ambiguity not attributable to uncertainties in calibration. Second, 17 percent is a relatively low error rate, given that each judgment is made without any prior experience with the voice and without the benefit of sentential context. Clearly there is a great deal of information within a single syllable which specifies the identity of its vowel nucleus. [Peterson and Barney (1952) report an even lower error rate, 5.6 percent, for 10 vowels in /h-d/ context with 10 talkers randomly mixed on each test.]

These data challenge the assumption that extended familiarization with a vowel space is the primary factor controlling vowel identification. The question

---

1 The talkers read the test syllables which were printed individually on cards. Standard English orthography served to represent seven of the syllables. In the two cases in which the target syllable was not a word, /pOp/ and /pUp/, the card specified "vowel as in cawed," and "vowel as in could." Most talkers pronounced the target syllables without hesitation and in most cases their tokens were recorded on magnetic tape after a single rehearsal. In no case did the experimenters provide a talker with spoken models.

2 Failures to respond were counted as errors. These occasions represent less than 2 percent of the total errors.
of what are the primary contributors must be reopened for study. In the following paper we report our studies of the information available within a single syllable. In the present paper we consider phonetic information that may extend across several syllables.

Because listeners' identification of vowels was better in the Blocked Condition than in the Mixed Condition, we may infer that information specifying the vowel must have been carried over the series of utterances of a single talker. We will first examine the advantage of keeping the talker constant on a vowel-by-vowel basis. Then we will test one hypothesis about the source of the information conveyed by talker constancy.

The errors made in identifying each intended vowel are shown in Figure 1; the columns indicate the percent of the time listeners made errors on each of the nine vowels. The hatched columns represent percent errors in the Mixed Condition, while the white columns represent errors in the Blocked Condition. For almost every vowel the percentage of errors drops in the Blocked Condition; the only exceptions are for /i/ and /a/. Three vowels, /i, r, u/, are readily identified in either condition. Of the remaining six vowels which are relatively ambiguous, only /a/ fails to show improvement, while familiarization definitely aids perception of /ɛ, æ, ç, â, ñ/.

But can we be sure that the improvements we observe are genuine? Shifts in response biases from one condition to the next could be responsible for some of these apparent improvements. A vowel could be correctly identified more often simply because it is more popular as a response. A direct sign of such a response bias is how often the vowel is used as an incorrect response to other vowels; when the vowel becomes more popular, the frequency of these false identifications increases.

The horizontal axis in Figure 2 indicates the change in correct identification between the Blocked and Mixed Conditions; placement to the right of the central vertical line represents superior performance on the Blocked Condition over that on the Mixed Condition. The vertical axis indicates the change in false identifications; placement above the central horizontal line represents greater frequency of false identifications on the Blocked Condition relative to the Mixed Condition.

True improvement may be defined as an increase in correct responses, coupled with a decrease in false identifications. Four of the more ambiguous vowels, /ɛ, æ, ç, ñ/, show genuine improvement by this measure. On the other hand, a change in correct identification that correlates with a change in false identification may be attributed to response biases alone. Thus the apparent improvement for /a/ may be attributed to a positive response bias, while /a/ shows a reciprocal negative bias.

These results demonstrate that familiarization with a talker's voice can yield genuine improvement in the identification of individual vowels. But what kind of information is available in a series of syllables? One common hypothesis is that tokens of several vowels are needed to specify accurately the shape of a talker's vowel space (Ladefoged and Broadbent, 1957). In fact, some authors have specifically suggested that the "point vowels" /i, â, u/ may be the primary calibrators of that space (Joos, 1948; Gerstman, 1968; Lieberman, Crelin, and
Figure 1: Mean percent errors in identification of each of nine vowels in /p-p/ environment.
Figure 2: Changes in correct identifications and biases in vowel perception attributable to keeping the talker constant throughout a test.
Klatt, 1972). The point vowels occupy the corners of the vowel triangle and are said to have the most determinate vocal tract area functions (Stevens, 1973).

If the point vowels do serve as calibration signals for normalization, two predictions follow directly. First, experience with a talker's point vowels should substantially reduce errors in identifying ambiguous vowels. Second, experience with another set of three vowels should not be as effective in reducing errors, or not be effective at all.

To make a direct test of these predictions we adapted our earlier Mixed Talker Condition by preceding each test syllable with a set of three vowels spoken by the same person. We used two sets of vowel precursors: /hi, ha, hu/ and /hr, hæ, hʌ/. The vowels were spoken in /h-/ syllables to facilitate articulation, while minimizing nonvocalic sources of information. If point vowels are the source of familiarization effects, the data in the /hi, ha, hu/ precursor condition should resemble that in the Blocked Condition of the earlier study.

Figure 3 displays the overall percent errors in the two precursor conditions, along with our earlier results: the 9.3 percent error rate in the Blocked Condition, the 17 percent error rate in the Mixed Condition without precursors, and the results for the /hi, ha, hu/ precursors and the /hr, hæ, hʌ/ precursors. The point vowel precursors improved identification only slightly, reducing errors from 17.0 to 15.2 percent; the difference is not statistically significant by a t-test. The three nonpoint vowels also reduced errors slightly, to 14.9 percent, though again the difference is not significant.

In other words, not only is there no evidence for a gain attributable to point vowels, but there is no difference between the point vowels and a set of nonpoint vowels. Overall, experience with specific sets of vowels seems to make little contribution to the total reduction of errors attributable to prior experience with a person's voice.

Before accepting these conclusions, it is worth checking whether there are improvements on specific vowels which are lost in the overall percentages. In Figure 4, the black columns indicate percent errors for each vowel in the Mixed Condition without precursors, the white columns are the results following the /hi, ha, hu/ precursors, and the hatched columns show errors following /hr, hæ, hʌ/. The point vowel precursors appear to help in identifying four vowels: /ɛ, ɔ, ʌ, ʊ/. Errors increase slightly for /æ/, and they increase on each of the precursor vowels, /i, a, u/. This may be a kind of contrast effect between tokens of the vowel in the precursor string and in the test syllable itself, i.e., subjects may be biased away from choosing a point vowel. If so, it suggests that the pattern of change observed with point vowel precursors may reflect merely a shift in response biases and not a real change in identifiability.

Figure 5 plots change in correct responses against change in false identification; each axis represents the difference between the Mixed Condition with point vowel precursors and the Mixed Condition without precursors. The four vowels /ɛ, ɔ, ʌ, ʊ/ that showed apparent improvement in Figure 4 all appear in the upper right-hand quadrant of Figure 5—i.e., all four reflect a positive bias and none shows true improvement. Two other vowels, /a, æ/, show a strong negative bias. A contrast effect with the precursor vowels is also evident, since each shows a negative response bias.
Figure 3: Mean percent errors in identification of nine vowels (averaged) in /p-p/ environment with and without precursors.
Figure 4: Mean percent errors in identification of each of nine vowels in /p-p/ environment with and without precursors.
Figure 5: Changes in correct identifications and biases in vowel perception attributable to /hi, ha, hu/ precursors.
It is clear that experience with a talker's point vowels does have an effect on subsequent identifications. But in no case do our data demonstrate genuine improvement in perceiving ambiguous vowels. The effect seems limited to a biasing of response probabilities.

Overall, these data challenge the notion that extended experience with a talker's voice is the primary source of information about his vowels; and in particular, they challenge the notion that the point vowels play a special role as calibrators of a presumed vowel space.

REFERENCES