# EVIDENCE OF TALKER-INDEPENDENT INFORMATION FOR VOWELS*

ROBERT R. VERBRUGGE**
and
BRAD RAKERD†
*Haskins Laboratories*
*and*
*The University of Connecticut*

The vowel information present in initial and final regions of /b/–vowel–/b/ syllables was examined in this study. Vowels were identified for unedited syllables spoken by a man and a woman, for the initial 20% of those syllables, for the final 20% of the syllables, for the initial and final 20% of the syllables combined and separated by a 60% silent gap, and for the initial and final 20% of the syllables interchanged across talkers and separated by a 60% silent gap. Results indicate: (1) that there is considerable vowel information present in the dynamic regions at the beginnings and endings of syllables; (2) that the information is, to a large extent, carried relationally by those regions; (3) that the information is talker-independent in form; and (4) that the information is complementary to, and distinct from, formant frequency information present in a syllable's center. An experiment assessing the perceived source(s) of these stimuli suggests that source perception is influenced by as yet unspecified acoustic modulations defined at the syllable level.

*Keywords:* acoustic targets, source perception, talker, vowel identification

## INTRODUCTION

When a vowel is coarticulated with preceding and following consonants to form a syllable, the resulting acoustic pattern usually includes periods of rapid spectral change at its beginning and end, and a period of relative spectral constancy at its center. It is well established that the configuration of formant frequency values present, or best approximated, at the syllable center provides information about the identity of the vowel (e.g., Joos, 1948; Ladefoged, 1975; Peterson and Barney, 1952). After Strange, Jenkins and Johnson (1983), we will refer to the ideal form of this configuration as an acoustic target.

There have been recurring indications that vowel information is also provided by the more dynamic regions of the syllable (Lehiste and Meltzer, 1973; Lindblom and Studdert-Kennedy, 1967; Shankweiler, Verbrugge and Studdert-Kennedy, 1978; Strange, Ver-

* This research was supported by Grants HD01994 and RR05596 to Haskins Laboratories. We are grateful to Winifred Strange, James Jenkins, and David Williams for their many helpful comments on this project.

** Now at AT&T Bell Laboratories.

† Now at Michigan State University.

brugge, Shankweiler and Edman, 1976). Perhaps the most compelling evidence of this comes from the experiments of Strange *et al.* (1983; also Jenkins, Strange and Edman, 1983). Those investigators assessed the perception of stimuli that preserved only the dynamic beginnings and ends of /b/–vowel–/b/ syllables, the syllable centers having been deleted and replaced with silence. Listeners spontaneously integrated the initial and final portions of these "silent-center" syllables, typically hearing a single utterance with an interruption in the middle (somewhat like a glottal stop). More importantly, vowel identification for these syllables was remarkably accurate, not differing significantly from the accuracy of identification for unedited syllables.

Two competing explanations for this silent-center finding provide the motivation for the present study. First, it is conceivable that listeners used the dynamic regions of those syllables to extrapolate to the formant-frequency targets which had been excised from the syllable centers. Lindblom (1963; also Lindblom and Studdert-Kennedy, 1967) has suggested that listeners make such extrapolations as a matter of course when processing natural speech. Whenever a talker speaks rapidly or destresses the production of a syllable, formant frequencies are "reduced," i.e., they fail to reach target values at the syllable center (Joos, 1948; Lindblom, 1963). Lindblom's (1963) proposal is that in these situations listeners draw on information in the dynamic regions to compute the missing targets. Specifically, they are said to draw on the fact that the initial and final formant trajectories form exponential functions which decelerate toward, or accelerate from, asymptotic target frequencies. To summarize, on this view the dynamic regions of a syllable contribute to vowel perception by subserving the more accurate estimation of target values approximated at the syllable center.

An alternative view of the silent-center result is that the dynamic regions convey vowel information that is complementary to, and distinct from, target information. One way to motivate this alternative is to think of vowels as articulatory events, that is, as gestures that manifest a characteristic organization of forces over the articulators (Fowler, 1977, 1980; Fowler, Rubin, Remez and Turvey, 1980). From this perspective, the vowels of a dialect are distinguished by different "styles" of articulatory movement. The resulting acoustic modulations provide substantial information about vowel identity, information that differs in kind from the target information present at a syllable's center.

To test the competing claims of the target-extraction and event-perception hypotheses, we constructed *hybrid* silent-center syllables, pairing the initial and final portions of corresponding syllables spoken by a man and a woman. According to the target hypothesis, a hybrid syllable should be very disruptive perceptually. Because the man and woman have different vocal tract sizes and shapes, their corresponding syllable portions should "point to" very different targets. This is illustrated in Figure 1. On the left are spectrograms of the man's and woman's productions of the syllable /bæb/. On the right those spectrograms have been cross-spliced to juxtapose their centers. It is clear that the center formant frequencies are quite discrepant, making it highly unlikely that any extrapolated target values could coincide across talkers.

According to the event hypothesis, a discrepancy in syllable centers is not necessarily disturbing. Talkers who speak a common dialect would be expected to produce a vowel with a common style of articulatory and acoustic change that is independent of idiosyn-

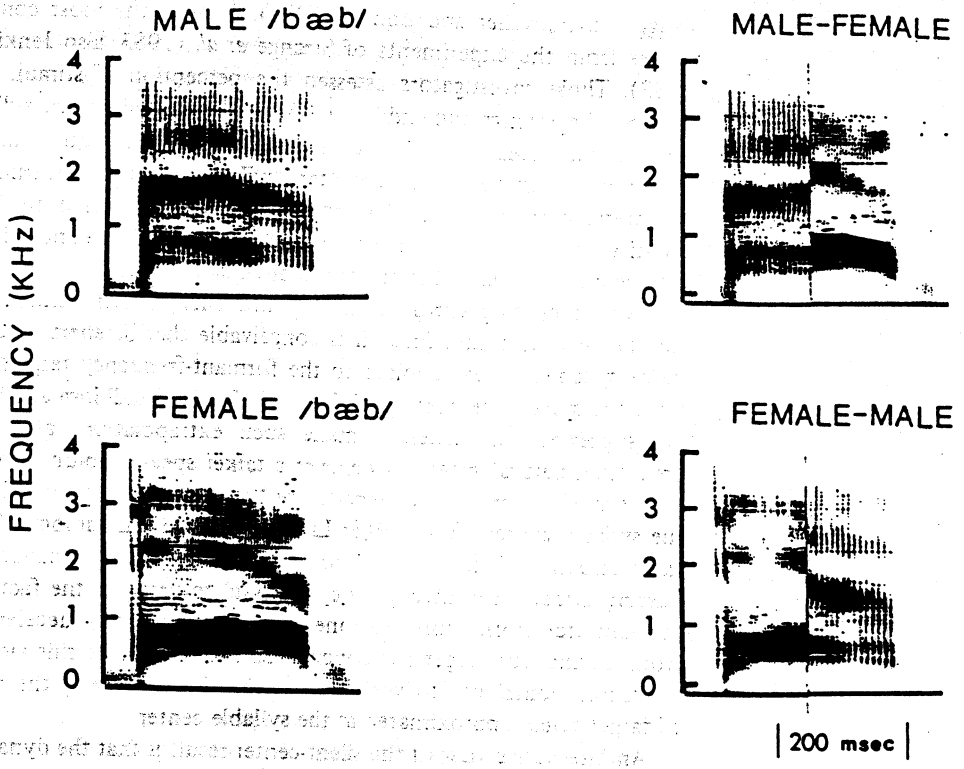


Fig. 1. Spectrograms of the man's and woman's productions of /bæb/ are presented on the left of the figure. To create the patterns on the right, those spectrograms were cut at the center of their voiced regions, and the initial and final halves were interchanged.

cratic differences in vocal tract size. Therefore, the event hypothesis, in its strongest form, predicts that the woman's and man's syllable portions should be integrated perceptually, and that accuracy of vowel identification should be high, perhaps as high as for single-talker silent-center syllables.

## Experiment 1: Vowel Perception

In this experiment we assessed the accuracy of vowel identification for hybrid-silent-center syllables and for a number of comparison syllables.

## Method

### *Stimuli*

The stimuli for all experimental conditions were derived from natural speech tokens of /b/–vowel–/b/ syllables. Syllable vowels were the American English vowels /i, ɪ, e, ɛ, æ, ɑ, ʌ, ɔ, o, ʊ, u/. A man and a woman each produced three tokens of each syllable. The syllables were produced in citation form and were paced to match the beat of a metronome. Productions were recorded on audio tape and then digitized for editing (sampling rate = 20 kHz). For each of the 11 vowels, we selected the pair of syllables, one from each talker, that were most closely matched in duration. In general it proved possible to find a very close match. The largest durational disparity was 20 msec and the average disparity was 4.5 msec (2% of the duration of the average voiced region, which was the same for both talkers).

*Spectral comparison 1: Between talkers.* The formants of the woman's vowels (*W* vowels) were typically higher in frequency than the formants of the corresponding vowels spoken by the man (*M* vowels). Table 1 reports their formant frequency values and shows that, on average, those values differed by 18%, 23%, and 17% for the first (*F*1), second (*F*2), and third (*F*3) formants respectively.[1] For comparison, we determined the average formant frequency differences between men and women based on Peterson and Barney's (1952) normative vowel data. That analysis is summarized in Table 2. Peterson and Barney found that formant values of an average adult female talker ($n = 28$) differed from those of an average male talker ($n = 33$) by 15% for *F*1, 17% for *F*2, and 17% for *F*3. The formant frequency differences between the two talkers of the present study were very close to these norms.

*Spectral comparison 2: Within talkers.* In absolute terms, the average formant frequency differences between our *W* and *M* vowels were 80 Hz for *F*1, 282 Hz for *F*2, and 420 Hz for *F*3. We wondered how these values compared with within-talker differences for the production of different vowels. Table 3 shows an analysis in which each talker's formant frequencies were rank-ordered and the differences between neighboring frequencies computed. The average differences were 24, 124, and 68 Hz respectively for *F*1, *F*2, and *F*3 of *M* vowels, and 40, 148, and 68 Hz for *F*1, *F*2, and *F*3 of *W* vowels. All of these values were less than half the size of between-talker production differences. We expect, therefore, that if a listener extrapolated to target values from the beginnings and endings of hybrid syllables, those targets would often be associated with *different* vowels.

The same expectation is supported by an analysis of the distribution of the two talkers' vowel tokens in *F*1–*F*2 space. Figure 2 shows that distribution, for a space in which the axes have been scaled to agree with those chosen by Peterson and Barney

[1] These figures are based on measurements of the nine vowels for which Peterson and Barney (1952) provide a comparison (/i, ɪ, ɛ, æ, ɑ, ʌ, ɔ, ʊ, u/). When we include in our analysis the vowels /e, o/, the woman's vowel formant frequencies differ from the man's by an average of 18%, 21%, and 17% for *F*1, *F*2, and *F*3, respectively.

TABLE 1

The woman's (W) and man's (M) formant frequency values in Hz, and their absolute differences expressed as a ratio of the man's values

| | First Formant | | | Second Formant | | | Third Formant | | |
|---|---|---|---|---|---|---|---|---|---|
| Vowel | W | M | (W–M)/M | W | M | (W–M)/M | W | M | (W–M)/M |
| i | 320 | 320 | 0.00 | 2480 | 2080 | 0.19 | 3240 | 2840 | 0.14 |
| ɪ | 400 | 480 | 0.17 | 2080 | 1760 | 0.18 | 2840 | 2480 | 0.15 |
| e | 320 | 400 | 0.20 | 2240 | 1920 | 0.17 | 3000 | 2560 | 0.17 |
| ɛ | 560 | 480 | 0.17 | 1840 | 1520 | 0.21 | 2560 | 2480 | 0.03 |
| æ | 640 | 560 | 0.14 | 2080 | 1480 | 0.41 | 2920 | 2480 | 0.18 |
| ɑ | 720 | 560 | 0.29 | 1320 | 1160 | 0.14 | 2920 | 2480 | 0.18 |
| ʌ | 640 | 480 | 0.33 | 1240 | 1080 | 0.15 | 3000 | 2480 | 0.21 |
| ɔ | 640 | 480 | 0.33 | 1240 | 1000 | 0.24 | 2920 | 2480 | 0.18 |
| o | 480 | 400 | 0.20 | 1000 | 920 | 0.09 | 2760 | 2320 | 0.19 |
| ʊ | 480 | 400 | 0.20 | 1160 | 1000 | 0.16 | 2760 | 2400 | 0.15 |
| u | 320 | 320 | 0.00 | 1160 | 840 | 0.38 | 2760 | 2160 | 0.28 |
| MEAN | | | 0.18 | | | 0.21 | | | 0.17 |
| /e, o/ excluded | | | 0.18 | | | 0.23 | | | 0.17 |

TABLE 2

Average women's (W) and men's (M) formant frequency values in Hz, and their absolute differences expressed as a ratio of the men's values. These data are from Peterson and Barney (1952)

| | First Formant | | | Second Formant | | | Third Formant | | |
|---|---|---|---|---|---|---|---|---|---|
| Vowel | W | M | (W–M)/M | W | M | (W–M)/M | W | M | (W–M)/M |
| i | 310 | 270 | 0.15 | 2790 | 2290 | 0.22 | 3310 | 3010 | 0.10 |
| ɪ | 430 | 390 | 0.10 | 2480 | 1990 | 0.25 | 3070 | 2550 | 0.20 |
| ɛ | 610 | 530 | 0.15 | 2330 | 1840 | 0.27 | 2990 | 2480 | 0.21 |
| æ | 860 | 660 | 0.30 | 2050 | 1720 | 0.19 | 2850 | 2410 | 0.18 |
| ɑ | 850 | 730 | 0.16 | 1220 | 1090 | 0.12 | 2810 | 2440 | 0.15 |
| ʌ | 760 | 640 | 0.17 | 1400 | 1190 | 0.18 | 2780 | 2390 | 0.16 |
| ɔ | 590 | 570 | 0.04 | 920 | 840 | 0.10 | 2710 | 2410 | 0.12 |
| ʊ | 470 | 440 | 0.07 | 1160 | 1020 | 0.14 | 2680 | 2240 | 0.20 |
| u | 370 | 300 | 0.23 | 950 | 870 | 0.09 | 2670 | 2240 | 0.19 |
| MEAN | | | 0.15 | | | 0.17 | | | 0.17 |

TABLE 3

The woman's and man's formant frequencies ($F$) in Hz, rank-ordered and differenced ($F$–$F$prev)

| Talker | First Formant | | Second Formant | | Third Formant | |
|---|---|---|---|---|---|---|
| | $F$ | $F$–$F$prev | $F$ | $F$–$F$prev | $F$ | $F$–$F$prev |
| woman | 320 | | 1000 | | 2560 | |
| | 320 | 0 | 1160 | 160 | 2760 | 200 |
| | 320 | 0 | 1160 | 0 | 2760 | 0 |
| | 400 | 80 | 1240 | 80 | 2760 | 0 |
| | 480 | 80 | 1240 | 0 | 2840 | 80 |
| | 480 | 0 | 1240 | 0 | 2920 | 80 |
| | 560 | 80 | 1840 | 600 | 2920 | 0 |
| | 640 | 80 | 2080 | 240 | 2920 | 0 |
| | 640 | 0 | 2080 | 0 | 2920 | 0 |
| | 640 | 0 | 2240 | 160 | 3000 | 80 |
| | 720 | 80 | 2480 | 240 | 3240 | 240 |
| | MEAN | 40 | | 148 | | 68 |
| man | 320 | | 840 | | 2160 | |
| | 320 | 0 | 920 | 80 | 2320 | 160 |
| | 400 | 80 | 1000 | 80 | 2400 | 80 |
| | 400 | 0 | 1000 | 0 | 2480 | 80 |
| | 400 | 0 | 1080 | 80 | 2480 | 0 |
| | 480 | 80 | 1160 | 80 | 2480 | 0 |
| | 480 | 0 | 1480 | 320 | 2480 | 0 |
| | 480 | 0 | 1520 | 40 | 2480 | 0 |
| | 480 | 0 | 1760 | 240 | 2480 | 0 |
| | 560 | 80 | 1920 | 160 | 2560 | 80 |
| | 560 | 0 | 2080 | 160 | 2840 | 280 |
| | MEAN | 24 | | 124 | | 68 |

(1952). Note that for 8 of the 11 vowel categories the man's token is closest to a token of a different vowel in the woman's space. In her case the mismatch is even more extreme; 10 of her 11 tokens lie nearest to a token of a different category in his space. This clearly indicates that the initial and final portions of a hybrid syllable would generally "point to" different target vowels when referred against a single talker's $F1$–$F2$ space.

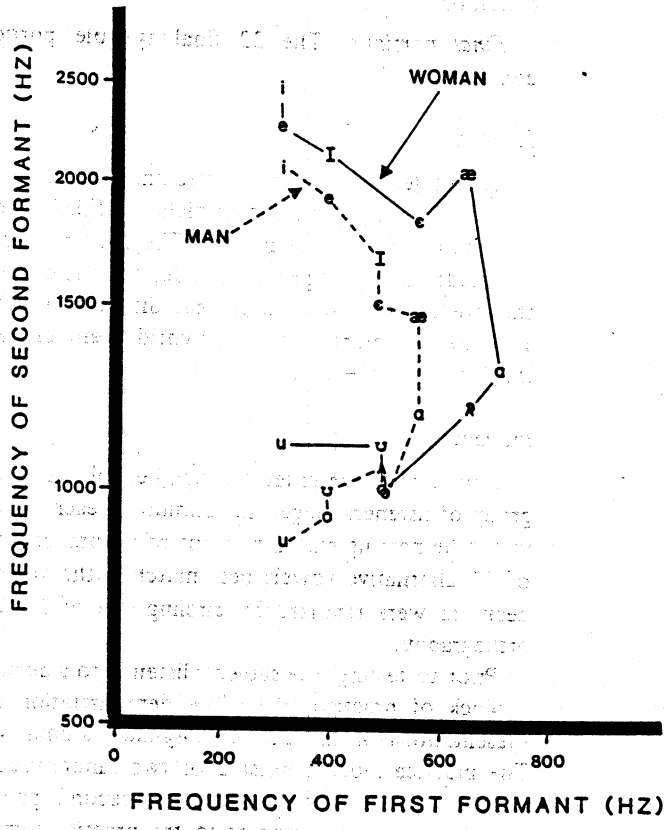


Fig. 2. Distribution of the man's and woman's vowels in an $F1/F2$ space.

*Experimental conditions*

The *W* and *M* syllables were edited for presentation in our experimental conditions according to the general procedures outlined by Strange *et al.* (1983). Each syllable was divided into three portions: (1) The initial portion of a syllable included the release burst of its initial /b/ plus 20% of the voiced region. (2) The central portion included the middle 60% of the voiced region. (3) The final portion included the final 20% of voicing plus the closure and release of the syllable-final /b/.[2] All measurements were made to the nearest zero-crossing of the speech waveform. Various combinations of the syllable

[2] Our editing procedures differed from those of Strange *et al.* (1983) and Jenkins *et al.* (1983) in terms of the percentage of the voiced region assigned to initial, center, and final syllable portions. Our choice of 60% as the center proportion is larger, on average, than their value, which varied from 50-60% depending on vowel category. As a result, our silent-center and hybrid-silent-center conditions involve a more severe deletion of signal.

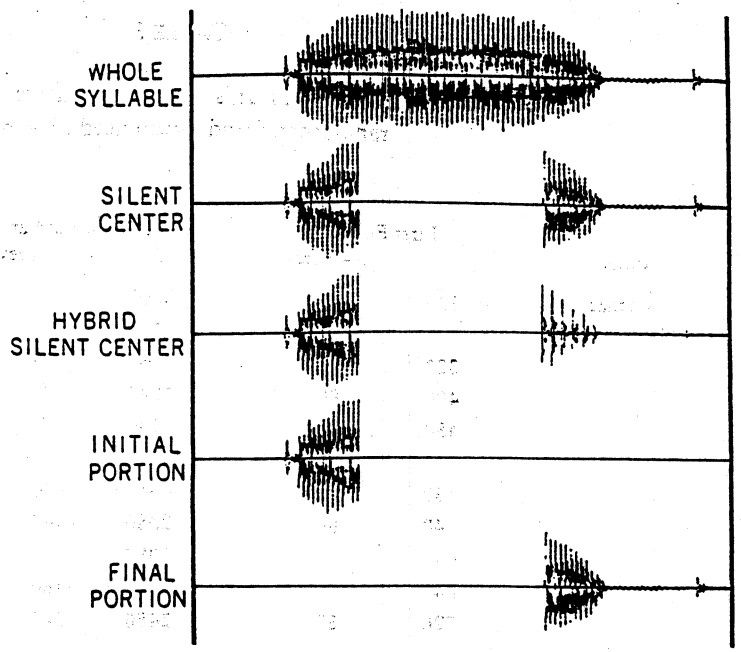


Fig. 3. Sample tokens of stimuli from the five experimental conditions as indicated. All stimuli derive from the woman's and man's productions of /bæb/.

portions were used to prepare the stimuli for five experimental conditions, as illustrated in Figure 3.

*Whole syllables.* For the whole-syllable condition, all three syllable portions were presented in their original temporal relation (i.e., the syllables were unedited). An example of a whole syllable, the woman's /bæb/, is shown at the top of Figure 3. There were 22 whole-syllable stimuli, 11 different syllables produced by each of the two talkers. These syllables are comparable to the "Control" syllables of Strange *et al.* (1983) and Jenkins *et al.* (1983).

*Silent centers.* Second from the top is an example of a silent-center syllable. The central portion of the woman's /bæb/ has been excised and replaced by silence in this instance. We created one silent-center version of each of the 22 syllables.

*Hybrid silent centers.* Third from the top of the figure is an example of a hybrid-silent-center syllable combining the initial portion of the woman's (*W*) /bæb/ with the final portion of the man's (*M*) /bæb/. The silent interval separating these portions was the same as for the woman's silent-center /bæb/. Eleven *W*/*M* and 11 *M*/*W* hybrids comprised the stimuli of this condition.

*Initial portions.* The 22 initial syllable portions provided the materials for this

condition.

*Final portions.* The 22 final syllable portions provided the materials for this condition.

*Subjects*

The subjects of this study were undergraduates enrolled in an introductory psychology course. Their participation partially fulfilled a course requirement. All of the subjects were native speakers of English. They had no known hearing difficulties and they had no knowledge of the hypotheses under test. The subjects were randomly assigned to one of the five experimental conditions, distributed as follows: whole-syllable condition ($n$ = 10), silent centers ($n$ = 15), hybrid silent centers ($n$ = 12), initial portions ($n$ = 11), final portions ($n$ = 11).

*Procedure*

Stimuli were presented through headphones at a comfortable listening level. A separate group of listeners judged the stimuli of each condition. The subjects were told that they would be hearing edited versions of natural speech, and that they were to decide which of 11 alternative vowels best matched the vowel that they heard on each trial. Their decisions were reported by circling one of 11 /b/–vowel–/b/ words written in English orthography.

Prior to testing, the subjects listened to a demonstration sequence and then completed a block of practice trials. The demonstration sequence consisted of two randomized presentations of the 22 whole-syllable stimuli with two-second pauses between them. The practice block consisted of two randomized presentations of the 22 stimuli of the condition to be tested, with four-second pauses between them. The subjects were required to make responses to the practice stimuli so that they would become familiar with the answer sheet; they were given no feedback as to the accuracy of those responses.

After the practice block, the subjects were allowed to ask questions of clarification about the testing procedure. The testing session commenced immediately after these questions. There were a total of 220 test trials, 10 randomized presentations of the 22 stimuli for a condition. A four-second pause separated succeeding stimuli. Subjects were given a five-minute break halfway through the test.

## Results and Discussion

The overall results for the five listening conditions are displayed in Figure 4. Each bar denotes the mean percentage of errors in vowel identification for the indicated condition, where an error was defined as a failure to categorize a vowel in the same way that the talker intended. Mean percentage errors by condition were as follows: whole syllables (8.8%), silent centers (23.1%), hybrid silent centers (27.4%), initial portions (56.4%), final portions (73.8%). Analysis of variance showed the differences in error rates across conditions to be highly significant [$F(4, 54) = 144.6; p < 0.001$]. Post hoc tests (Newman-Keuls) revealed that all pairwise differences among the conditions were

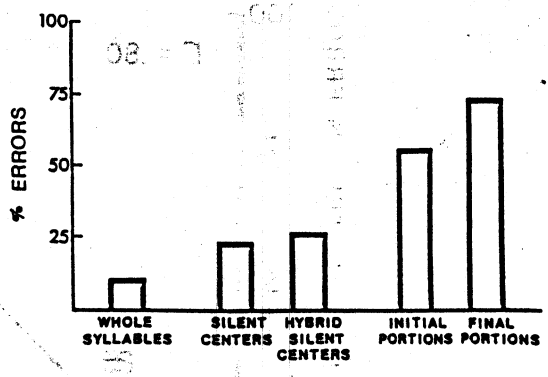


Fig. 4. Vowel identification error rates for the five experimental conditions. Errors are pooled over 11 vowels and over two talkers.

significant ($p < 0.01$) with one exception: There was no statistically significant difference between the silent-center and hybrid-silent-center conditons ($p > 0.05$).

*Comparison with previous silent-center studies*

Our results replicate and extend the central finding of previous studies examining silent-center stimuli (Jenkins *et al.*, 1983; Strange *et al.*, 1983) – namely, that subjects can identify vowels with good accuracy when syllable centers are silenced. Our results also replicate the previous finding that vowel perception is poor when either initial syllable portions or final portions are presented alone. These results imply, on the one hand, that the dynamic beginnings and endings of syllables are a rich source of information about the syllable vowel and, on the other, that the information is somehow conveyed *relationally* by those beginnings and endings.

One contrast with past studies is our observation of a significant difference between the silent-center and whole-syllable conditions. Previous investigators found no differences between these two conditions (Jenkins *et al.*, 1983; Strange *et al.*, 1983). We may have found a difference in this study because, on average, we deleted a somewhat greater portion of the signal in our silent-center condition than was deleted by others (see footnote 2). Other possible explanations are that there were between-study differences in familiarization with the materials, or in other aspects of the training, or in the subject populations themselves. The overall error rates for both our whole-syllable and silent-center conditions were higher than those seen in previous studies, indicating that others were operating much nearer to the error "floor."

*Silent centers vs. hybrid silent centers*

Of greatest interest to us was the finding that the hybrid and silent-center conditions did not differ significantly. This strongly suggests that the vowel information preserved

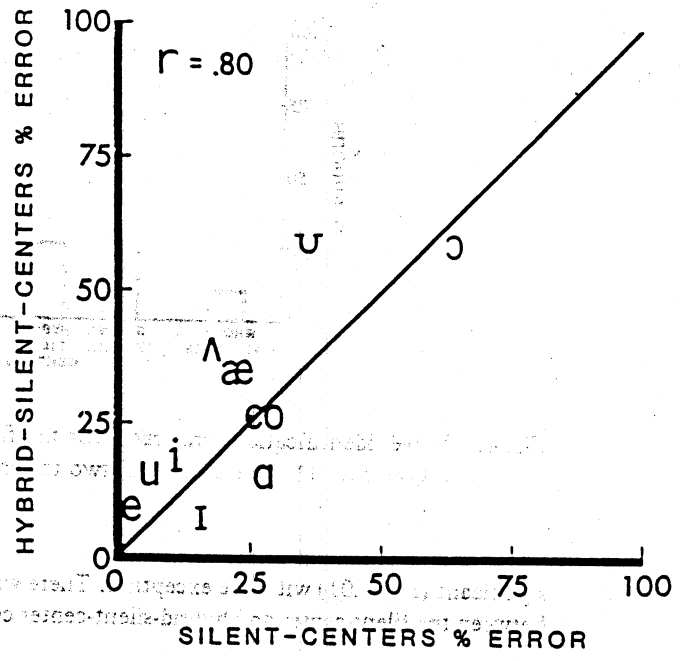


Fig. 5. Scatter plot of errors for 11 vowels presented in silent-center (abscissa) and hybrid-silent-center (ordinate) conditions. Silent-center errors are collapsed across two talkers, hybrid-silent-center errors are collapsed across man-woman and woman-man hybrids. The coefficient of correlation ($r$) is also provided.

in silent-center syllables is also preserved in hybrids, despite their change of source. That suggestion is strengthened further by a vowel-by-vowel comparison of errors made in the silent-center and hybrid conditions. The comparison is illustrated in Figure 5. Plotted on the abscissa are errors for the silent-center syllables (collapsed over talkers) and on the ordinate are errors for the hybrid syllables (collapsed over *M/W* and *W/M* versions). The two data sets are highly correlated ($r = 0.80$; $p < 0.01$), and the clustering of points about the diagonal of the figure demonstrates how similar the errors are in absolute terms. The implication of all of these results is that the vowel information in dynamic regions of a syllable is largely invariant across talkers. It is highly unlikely that this dynamic information subserves the perceptual extraction of any sort of acoustic target, since targets are highly variant across talkers. It is much more likely that the information is indicative of a characteristic articulatory style that is common to productions of the same vowel by talkers of the same dialect.

*The role of syllable duration*

Following others (Jenkins *et al.*, 1983; Strange *et al.*, 1983), we have proposed that

the dynamic information for vowels is carried relationally by the initial and final syllable portions. Perhaps the simplest relation that might carry it is a durational one. One could imagine that information about the duration of the syllable as a whole could help a listener to distinguish between spectrally-similar, durationally-different vowels in the syllable nucleus. Two lines of evidence speak against this hypothesis. The first comes from a previous study (Strange *et al.*, 1983) that included conditions in which durational differences among silent-center syllables were neutralized. In one condition all of the silent intervals were set equal to the shortest silent duration in the test set and in another they were set equal to the longest. Neither manipulation significantly affected the outcome when all stimuli were produced by a single talker. The "lengthening" manipulation did produce a small but significant increase in errors when different talkers' syllables were interspersed; however, this increase was manifest for vowels of all categories, not just for the short vowels, suggesting that factors other than vowel duration were affected. Overall, there was very little evidence that durational differences among silent-center syllables are an important source of vowel information.

Very little evidence of this can be found in our own results as well. If duration were a primary factor, one would expect the lower error rates for silent centers and hybrids, relative to the initial and final syllable portions, to be due primarily to a reduction in short-long vowel confusions. Short-long vowel errors would be high for the isolated portions (where only spectral information is available), and low for the silent centers, because these syllables presumably supply the duration information needed to distinguish between spectrally-similar short and long vowels.

The first row of Table 4 provides a summary of errors for four spectrally-similar, durationally-different pairs of monophthongs, for each condition of Experiment 1. The second row of the table presents overall errors for the eight vowels after short-long confusions have been removed. The third row summarizes the errors specifically due to short-long confusions. With respect to the duration hypothesis, two observations seem important. First, by the strong form of this hypothesis, errors on isolated portions are due primarily to short-long ("duration") confusions, and overall errors should therefore be roughly *equal* for silent centers and for the isolated portions after duration errors have been removed. The data in Table 4 (second row) do not support this prediction. Second, while more duration errors are observed for isolated portions than for silent centers (third row), the *proportion* of errors attributable to short-long confusions stays relatively constant across these conditions (see fourth row of the table). This suggests that the silent-center format *does not* differentially reduce duration-based errors, but has a broader, and different, kind of impact in reducing perceptual errors. Parametric studies using a broader set of stimulus materials will be needed to address this question further.

*Modeling the relationship between initial and final portions*

If the initial and final syllable portions are not affording listeners a better estimate of intrinsic vowel duration, then how is it that perception is so much better in the silent-center and hybrid-silent-center conditions? One might argue that it is better because in

TABLE 4

Mean percentage errors on eight vowels, /ɪ, i, ɛ, æ, ʌ, ɑ, ʊ, u/, including and excluding confusions on adjacent short-long vowels

| | Whole Syllable | Silent Center | Hybrid Silent Center | Initial Portion | Final Portion |
|---|---|---|---|---|---|
| Overall errors | 8.3 | 20.1 | 26.3 | 47.9 | 65.9 |
| Overall errors, excluding short-long errors[a] | 3.5 | 11.8 | 17.8 | 26.3 | 39.7 |
| Short-long errors | 4.8 | 8.3 | 8.5 | 21.6 | 26.2 |
| Proportion[b] | 0.58 | 0.41 | 0.32 | 0.45 | 0.40 |

[a]Short-long errors are confusions within any one of the following four vowel pairs: /ɪ–i/, /ɛ–æ/, /ʌ–ɑ/, /ʊ–u/.

[b]Short-long vowel errors as a proportion of overall errors.

these conditions listeners are, in effect, given two chances to identify the vowel, one chance based on the initial portion and a second based on the final portion. In this section we consider this alternative.

How might initial-portion and final-portion percepts be processed to derive a single vowel judgment? The simplest possibility is that those percepts are, for each vowel, perfectly independent and that a listener simply chooses between them at random. If so, we would expect that errors in the silent-center and hybrid conditions should average 71% (the mean of initial- and final-portion error rates). A nonrandom selection process could, at best, produce errors rates of 56% (taking the better of the initial- and final-portion rates for each vowel). Even the latter prediction is much higher than the actual error rates observed for silent centers (23%) and hybrids (27%). Moreover, it poorly predicts the ordering of error rates across vowels: The correlation between the nonrandom guessing prediction and the observed errors for silent centers was 0.41, and for hybrids it was 0.42.

One might propose a more sophisticated decision model in which the initial- and final-portion percepts are processed in contingent fashion to arrive at a vowel response. For example, the initial portion could be used to narrow down the set of alternatives and the final portion to make a selection from among this reduced set. A good candidate for the initial classification is the intersection of two major phonetic dimensions: high-vs.-low and front-vs.-back. With respect to this four-way classification, listeners made an average

of 26% errors when categorizing vowels in the initial-portion condition (excluding the diphthongs /e, o/). Estimates of the probability for error when making the final selection within these categories can be derived from our data on the final portions. When the probabilities for error in the two stages are combined, one obtains a predicted error rate for judgments on the silent-center and hybrid syllables as a whole.[3] Figure 6 shows the comparison between predicted and observed errors for the hybrid condition (the silent-center comparison looks similar). Like the previous models, this contingent model generally overpredicts the absolute level of errors and poorly predicts the patterning of errors among the vowels. The correlations between predicted and observed errors were 0.27 for hybrid vowels and 0.50 for silent-center vowels.

The models we tested all assumed that the syllable portions were analyzed separately, and were only related at a late stage in a decision process. This type of perceptual analysis would seem to be demanded by the target-extraction view, which proposes that on-glide and off-glide functions separately specify a target. In particular, separate perceptual analyses would seem to be the only way the target view could approach the perception of *hybrid* syllables, since the syllable portions specify very different asymptotic targets in this case. However, all of the "separate analysis" models underpredict listeners' accuracy on the hybrids by a wide margin. This strongly suggests that a listener does not process the syllable portions separately but, instead, derives vowel information from a "superadditive" relation between them. In other words, it suggests that some singular function over the two portions of a hybrid is detected by the perceiver as the basis for a vowel judgment.

This account of the hybrid-syllable results is compatible with the event hypothesis, which holds that the early and late stages of an event should bear a principled relation to one another. Defining such relations in acoustic terms is a major challenge for future research. The simplest possibility is to define a duration measure over the hybrid syllable as a whole. However, neither our results nor those of Strange *et al.* (1983) provide much support for syllable duration as the critical "superadditive" relation (see previous section). More complicated possibilities involve characteristic frequency and amplitude modulations over the syllable. Whatever function we may discover, our hybrid data suggest that it will be talker-independent in form, and that it will not be the sum of two exponentials sharing a common asymptote.

The judgment models also raise questions about the role of more local sources of information for vowel identity. The contingent model, for example, considered the possibility that different regions of a syllable provide different kinds of information. While that particular model proved uninformative, there was some evidence in listeners' errors on the isolated portions that the early and late regions of a syllable carry some information about vowel properties. Listeners in both conditions showed better-than-chance performance (chance would be 91% errors). Also, as we noted above, the initial

[3] For example, in the final-portion condition the high-back vowels /u/ and /u/ were confused with one another on 8% (/u-u/) and 34% (/u-u/) of trials. These percentages, in combination with the probability of making an error when categorizing a high-back vowel as high-back in the initial-portion condition (20%), provided our contingent estimates for /u/ and /u/.

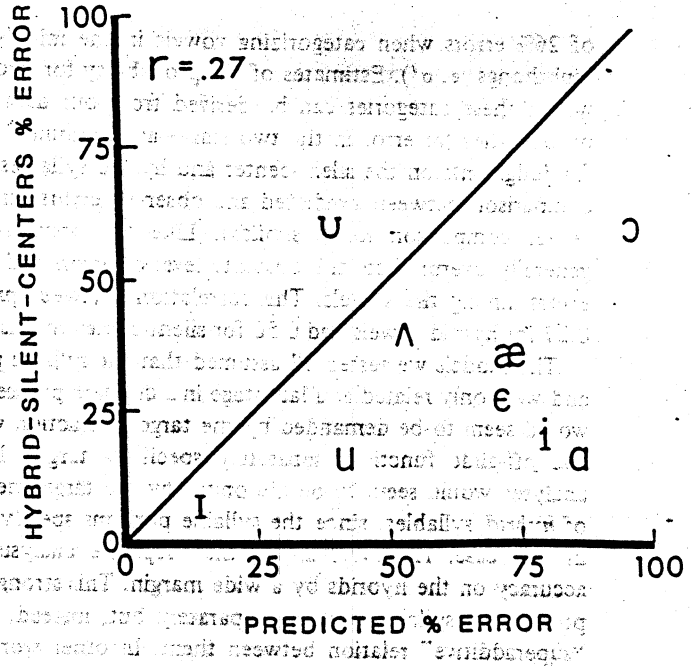


Fig. 6. Scatter-plot (and correlation) of errors on hybrid silent-center syllables, as predicted by a contingent-judgment model (abscissa) and as observed in the identification test (ordinate).

portions of the monophthongs carried sufficient information to support four-way classification (high-low, front-back) with only 26% errors. A similar analysis of errors on final portions shows 33% errors for the four-way classification.

These results on the isolated portions raise a second challenge for future research: to identify the carriers of information in these more local regions of a syllable. In the case of the initial portions, one candidate is the release burst of initial stop consonants. In fact, several studies have reported that this brief initial phase of a syllable is sufficient for better-than-chance discrimination within small sets of vowels (Blumstein and Stevens, 1980; Winitz, Scheib and Reeds, 1972). The acoustic basis for these effects is still not clear, nor is it clear how well listeners could do on a larger, more representative set of vowels. Even so, these findings provide a good example of a general principle we seek to develop in this paper: The transient regions of a syllable may provide information that is specific to a vowel without necessarily being information about a target state. A rough analogy can be drawn to the role of onset transients in the identification of musical instruments. The dynamic structure of these transients carries more information about instrument identity than does the steady-state region of a sustained tone (Grey and Gordon, 1978; Luce and Clark, 1967; Saldanha and Corso, 1964). More to the point, the transients do not simply aid the extraction of steady-state timbre; they provide

information which is different in kind. In the case of vowels, we expect to find a similar pattern: namely, that the structure of a talker's onset transients is both specific to the vowel and distinct from spectral targets.

## EXPERIMENT 2: SOURCE PERCEPTION

After the completion of each session of testing in Experiment 1, we informally interviewed subjects about their impressions of the edited syllables and were surprised to discover that subjects in the hybrid condition rarely heard a complete change of source. Instead, they heard a single talker, typically a male, and, more particularly, a male prone to abrupt pitch changes. These reports were surprising because the hybrid stimuli contain marked discontinuities of fundamental and formant frequencies, and these would normally be expected to specify a change of articulatory source. The perceptual reports suggest that the hybrid syllables contain other types of acoustic information, which strongly specify a single production by a single source. In the normal course of events, this acoustic structure would parallel other information about the source, such as fundamental frequency and formant frequency contours. However, in the unique case of the hybrid stimuli, it opposes these other sources of information and appears to predominate over them. Since this speculation has implications for the study of source perception, we thought it important to make a more rigorous test of the findings that prompted it. In Experiment 2, we directly sought subjects' judgments of the number of talkers they heard when listening to hybrid silent-center (and silent-center) stimuli.

### Method

#### *Subjects*

Nine undergraduate students were the subjects of this experiment. They were native speakers of English with normal hearing. They had had no contact with the subjects of Experiment 1 and were not themselves subjects of that experiment.

#### *Stimuli*

The stimuli of this experiment were the silent-center and hybrid-silent-center stimuli described in Experiment 1.

#### *Procedure*

Ten randomized repetitions of the 22 silent-center stimuli, spaced at four-second intervals, comprised a silent-center test block. A comparable arrangement of the 22 hybrid stimuli comprised a hybrid test block. Each test block was presented to subjects twice, in alternation. Five subjects began with the silent-center block, four began with the hybrid block. The subjects' task was to determine which of the following three alternatives best described the source(s) of the stimuli heard on each trial:

(1) One talker speaking with normal intonation.

(2) One talker speaking with a pitch change.
(3) Two talkers speaking.

Responses were reported by checking off the appropriate alternative on an answer sheet.

Prior to testing, subjects completed a practice block in which they responded to one presentation of each hybrid and silent-center stimulus. The order of these presentations was randomized. The subjects received no feedback regarding the accuracy of their responses. The practice block was followed by a pause for questions regarding procedure, and then by the first test block. There was a five-minute break between test blocks. All testing was completed in a single session.

## Results and Discussion

The results of this experiment are summarized in Figure 7, which shows the proportion of silent-center and hybrid responses in each category (collapsed across the two orders of presentation). The results confirm the informal reports given by subjects in the hybrid condition of Experiment 1: Hybrid stimuli are most generally perceived to have been produced by a single talker. They were so perceived on a total of 75% of the trials in the present experiment. That percentage was only slightly smaller than the total percentage of single-talker responses for silent-center syllables (82%). The principal difference between hybrid and silent-center responses was in their distribution over the two single-talker categories. With silent-center stimuli, subjects more often judged that the talker spoke with normal intonation (57% of all judgments), while with hybrid stimuli, subjects more often heard a pitch change (43% of all judgments).

Listeners' judgments that the hybrid stimuli derived from a single source may have been facilitated by the presence of the silent gap between the initial and final portions spoken by the different talkers. The stimuli did not contain *instantaneous* changes in fundamental frequency and formant contours. Instead, those contours were heard to be interrupted at one point and resumed at another. Perhaps in such cases it is reasonable for listeners to ascribe the gap's "bridge" to the rather curious behavior of a single talker. If so, we would note that there is a strong asymmetry in those ascriptions. With both woman/man and man/woman hybrids, listeners nearly always reported that the single talker was a man. For some reason his vocal characteristics predominated.

We would also note that few perceptual gaps can be "bridged" so readily as the hybrid gap. While a pitch break of the magnitude seen across the initial and final portions is conceivable for a single talker, a formant pattern break of the magnitude seen (15-20%) is inconceivable. (It would require a change in the talker's age or sex in mid-utterance.) Listeners integrated the syllable portions in spite of this radical change in effective vocal tract dimensions, and this suggests that other, more powerful information for source continuity was present in the acoustic signal. It seems likely that listeners were strongly aided in bridging the silent gap by the common style with which the two talkers produced the original syllables. The two talkers spoke the same dialect and produced the same vowel gestures, in the same phonetic context, under the same timing regimen (matching the beats of a metronome). The close similarity of their articulatory styles would

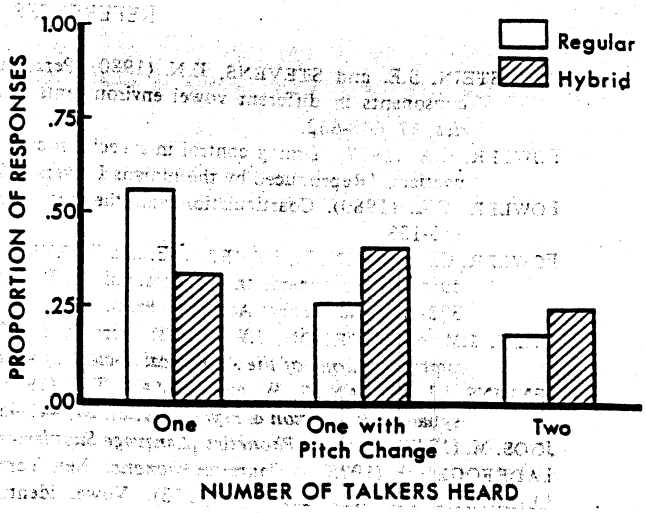


Fig. 7. Proportion of trials on which subjects judged regular (single-talker) and hybrid silent-center stimuli to have been produced by: (1) one talker speaking with a normal pitch; (2) one talker speaking with an abrupt pitch change; or (3) two talkers.

produce, as a natural consequence, a close similarity of acoustic "styles of change" in their productions. These dynamic consequences of "producing the same vowel with the same timing" may be the basis for subjects' integrating the two portions perceptually and hearing them as the product of a common source. Given the composition of the hybrid syllables, we can conclude that this acoustic information is defined over the syllable as a whole, and, in particular, that it is defined sufficiently by a relation between the initial and final regions of the syllable.

## CONCLUSION

Experiments 1 and 2 provide strong indications that the perception of vowel identity and source continuity is sensitive to dynamic acoustic structure defined over the course of a whole syllable. The acoustic information appears to be distinct in type from such variables as syllable duration and spectral targets (whether realized in the signal or extrapolated). Vowel perception and source perception can be remarkably impervious to discontinuities in local spectrum, if speech materials are otherwise matched in timing and articulatory style. This strongly suggests that a dialect's vowels can be characterized by higher-order variables (patterns of articulatory and spectral *change*) that are independent of a specific talker's vocal tract dimensions. A more precise definition of these variables will aid our understanding of the acoustic basis for identifying a vowel and, not coincidentally, for perceiving an articulation as continuous.

## REFERENCES


BLUMSTEIN, S.E. and STEVENS, K.N. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America*, 67, 648-662.

FOWLER, C.A. (1977). Timing control in speech production. Ph.D. dissertation, University of Connecticut. [Reproduced by the Indiana University Linguistics Club, Bloomington, Indiana.]

FOWLER, C.A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, 8, 113-133.

FOWLER, C.A., RUBIN, P., REMEZ, R.E. and TURVEY, M.T. (1980). Implications for speech production of a general theory of action. In B. Butterworth (ed.), *Language Production* (pp. 373-420). New York: Academic Press.

GREY, J.M. and GORDON, J.W. (1978). Perceptual effects of spectral modifications on musical timbres. *Journal of the Acoustical Society of America*, 63, 1493-1500.

JENKINS, J.J., STRANGE, W. and EDMAN, T.R. (1983). Identification of vowels in "vowelless" syllables. *Perception & Psychophysics*, 34, 441-450.

JOOS, M. (1948). *Acoustic Phonetics* [*Language Supplement*, 24].

LADEFOGED, P. (1975). *A Course in Phonetics*. New York: Harcourt Brace Jovanovich.

LEHISTE, I. and MELTZER, D. (1973). Vowel identification in natural and synthetic speech. *Language and Speech*, 16, 356-364.

LINDBLOM, B.E.F. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, 35, 1773-1781.

LINDBLOM, B.E.F. and STUDDERT-KENNEDY, M. (1967). On the role of formant transitions in vowel recognition. *Journal of the Acoustical Society of America*, 42, 830-843.

LUCE, D. and CLARK, M. Jr. (1967). Physical correlates of brass-instrument tones. *Journal of the Acoustical Society of America*, 42, 1232-1243.

PETERSON, G.E. and BARNEY, H.L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24, 175-184.

SALDANHA, E.L. and CORSO, J.F. (1964). Timbre cues and the identification of musical instruments. *Journal of the Acoustical Society of America*, 36, 2021-2026.

SHANKWEILER, D.P., VERBRUGGE, R.R. and STUDDERT-KENNEDY, M. (1978). Insufficiency of the target for vowel perception. *Status Report on Speech Research* (Haskins Laboratories, New Haven, Conn.), SR-55/56, 103-111.

STRANGE, W., VERBRUGGE, R.R., SHANKWEILER, D.P. and EDMAN, T. (1976). Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America*, 60, 213-224.

STRANGE, W., JENKINS, J.J. and JOHNSON, T. (1983). Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America*, 74, 694-705.

WINITZ, H., SCHEIB, M.E. and REEDS, J.A. (1972). Identification of stops and vowels for the burst portion of /p, t, k/ isolated from conversational speech. *Journal of the Acoustical Society of America*, 51, 1309-1317.