Speech Communication 6 (1987) 1–14
North-Holland

CATEGORICAL TENDENCIES IN IMITATING SELF-PRODUCED ISOLATED VOWELS

Bruno H. REPP and David R. WILLIAMS*
Haskins Laboratories, 270 Crown Street, New Haven, CT 06511–6695, U.S.A.

Received 24 February 1986
Revised 2 September 1986

Abstract. An earlier experiment requiring literal imitation of synthetic isolated vowels from [u]–[i] and [i]–[æ] continua [8] was replicated using as stimuli vowels produced by the subjects themselves. Even though imitation accuracy was much improved, the responses deviated from the stimuli in ways similar to those observed previously with synthetic stimuli. That is, categorical tendencies (nonlinear stimulus-response mappings of formant frequencies, nonuniform response variability across each continuum, and peaks in formant frequency distributions) were obtained even with stimuli that matched the subjects' articulatory capabilities. This rules out one possible explanation of the observed categorical tendencies, viz., that they arise in the perceptual translation of synthetic stimuli into a talker's production space.

Zusammenfassung. Ein früheres Experiment, in dem isolierte synthetische Vokale van [u]–[i] und [i]–[æ] [8] Kontinuen zu imitieren waren, wurde wiederholt mit Vokalen, die von den Versuchspersonen selbst gesprochen waren. Obwohl die Genauigkeit der Imitation nun viel größer war, wichen die Imitationen dennoch von den Stimuli ab, in ähnlicher Weise wie früher von den synthetischen Stimuli. Das heißt, wir fanden kategorielle Tendenzen (nonlineare Beziehungen zwischen den Formantfrequenzen der Stimuli und der Imitationen, ungleiche Variabilität der Imitationen über jedes Kontinuum, und Gipfel in der statistischen Verteilung der Formantfrequenzen) mit Stimuli, die den artikulatorischen Möglichkeiten der Versuchspersonen genau angepaßt waren. Dieses Resultat schließt eine mögliche Erklärung der beobachteten kategoriellen Tendenzen aus—nämlich, daß sie während der perzeptorischen Übersetzung von synthetischen Stimuli in den artikulatorischen Bereich eines Sprechers entstehen.

Résumé. Une expérience antérieure réclamant l'imitation exacte de voyelles synthétiques isolées sur des continuum [u]–[i] et [i]–[æ] [8] a été rédupliquée à l'aide de stimuli vocaliques produits cette fois par les subjets eux-mêmes. Bien que la précision de l'imitation ait été nettement améliorée, les réponses obtenues dévient des stimuli de manière similaire à ce qui avait été observé avec les stimuli synthétiques. C'est-à-dire que des tendances catégorielles (projection non-linéaire des fréquences formantiques du domaine des stimuli vers le domaine des réponses, variabilité non-uniforme des réponses au travers de chaque continuum, maxima dans la distribution des fréquences formantiques) ont été obtenues même avec des stimuli qui s'accordaient aux capacités articulatoires des sujets. Ce résultat écarte l'idée que les tendances catégorielles observées puissent s'expliquer par la transposition perceptive des stimuli synthétiques vers l'espace articulatoire du locuteur.

Keywords. Categorical perception, imitation of self-produced vowels, isolated synthetic vowels, production space.

1. Introduction

In a recent study [8], we investigated the claim [1, 4] that subjects' vocal imitations of isolated, steady-state vowels follow a categorical pattern. Two subjects (the authors) imitated synthetic vowels from 12-member [u]–[i] and [i]–[æ] continua at three different temporal delays, which had little effect on response patterns. The functions relating stimulus and (average) response formant frequencies across each vowel continuum exhibited local changes in slope, response standard deviations varied, and the distributions of response formant frequencies showed distinct peaks and valleys. The response patterns thus showed cate-

* Also at Department of Psychology, University of Connecticut, Storrs, CT 06268, U.S.A.

gorical tendencies, but few instances of strictly categorical responses (i.e., identical responses to different stimuli representing the same vowel category).

Where do these categorical tendencies in imitation come from? There are at least four independent (but not mutually exclusive) possibilities, some perhaps more plausible than others. The tendencies could originate either in the subjects' *perception* of the stimulus vowels or in their *production* of the imitations. On the perceptual side, there are two possibilities:

(1) Perceptual nonlinearities might arise when the stimuli are synthetic and/or not well matched to the subject's production capabilities. An additional stage of translation may be required between such stimuli and the vocal response, and certain irregularities could arise at that stage.

(2) Phonetic categorization may intrude upon the internal representations of the stimuli, as it apparently does in vowel discrimination tasks [6, 7]. In other words, the imitation task may simply elicit the same quasi-categorical response pattern that is typically obtained in vowel experiments following the "categorical perception" paradigm.

On the production side, there are two additional possibilities:

(3) The observed stimulus-response nonlinearities may reflect articulatory constraints on vowel production that are either universal or acquired through experience with a particular language (notwithstanding the relative rarity of isolated vowels in everyday communication). This hypothesis was favored by Chistovich et al. [1].

(4) Finally, there is the possibility that the constraints are not articulatory but acoustic in nature, in that certain discontinuities in the transform from vocal tract shape to the output lead a speaker to favor certain formant patterns, as suggested by Stevens' "quantal theory" of vowel production [9].

The first hypothesis seems perhaps less plausible than the others in view of the fact that Chistovich et al. [1], in their original demonstration of categorical imitation, used synthetic stimuli that were modelled after the (single) subject's own productions. On the other hand, that hypothesis is the easiest one to test and deserves to be ruled out before the other possibilities are investigated more thoroughly. This was the purpose of the present study.

Acoustic analysis of the responses obtained in our earlier study [8] revealed a large variety of formant patterns, which made it possible to select a number of utterances that formed naturally produced vowel continua specific to each subject. With this assurance that each subject was physically able to produce a precise match for each stimulus, we proceeded to replicate the experiment. Subjects, design, and procedure were identical, and the reader is referred to our earlier report [8] for some methodological details and for results not reproduced here. (Figs. 1–8 correspond to earlier figures with the same numbers.)

Even though there were only two subjects in this study (due to our method of stimulus selection, our desire to make a within-subject comparison with the earlier results, and our preference for experienced subjects), we expected to have sufficient evidence against the hypothesis under test if (1) each subject's imitation responses along a vowel continuum showed significant, nonuniform deviations from the stimulus parameters, and (2) these deviations followed a pattern similar to that obtained in our earlier study.

2. Methods

2.1. Stimuli

Two 12-member vowel continua, one intended to range from [u] to [i] and the other from [i] to [æ], were selected from appropriate two-dimensional scatter plots of each subject's imitation responses in the first study [8]. Each formant frequency plot included 36 responses to each of 12 members of a synthetic vowel continuum, either [u]–[i] or [i]–[æ], a total of 432 data points. From each of these plots we selected twelve tokens that were as equidistant as possible and followed a pre-determined path in the (linearly scaled) formant frequency space. The resulting natural [i]–[æ] continuum was selected to fall along a straight line in the F_1-F_2 plane, determined by linear regression of F_2 on F_1 in the scatterplot, whereas the [u]–[i] continuum was made to follow a curve

in the F_2-F_3 plane, derived by eye from the central tendencies in the data. In addition, since it was not possible to vary other stimulus parameters systematically, it was attempted to hold F_1 on the [u]–[i] continuum, and F_3 on the [i]–[æ] continuum, as constant as possible by avoiding tokens with deviant values. Extreme values of fundamental frequency and duration were likewise excluded by listening to each continuum and by replacing tokens that "stuck out." The average formant frequencies of the stimuli selected, determined by LPC analysis, are listed in Table 1. Stimulus durations varied between 150 and 210 ms, average fundamental frequencies between 104 and 127 Hz (DW) and between 111 and 132 Hz (BR)[1].

Table 1
Average formant frequencies of stimulus vowels (Hz)

	DW			BR		
Stim	F_1	F_2	F_3	F_1	F_2	F_3
[u]–[i] continuum						
1	310	1036	2071	310	979	2068
2	301	1164	2102	312	1101	2025
3	308	1296	2122	318	1232	1986
4	305	1431	2135	309	1330	1995
5	308	1538	2154	320	1466	2030
6	308	1620	2210	312	1564	2075
7	307	1694	2308	319	1673	2141
8	313	1778	2378	324	1782	2195
9	312	1858	2453	317	1877	2257
10	307	1917	2523	309	1966	2359
11	302	2022	2617	297	1996	2451
12	276	2089	2666	293	2027	2568
[i]–[æ] continuum						
1	269	2124	2629	297	2069	2592
2	300	2082	2488	313	2038	2533
3	334	2035	2442	341	2014	2512
4	370	2001	2457	366	1985	2460
5	381	1963	2453	383	1934	2378
6	414	1911	2396	412	1895	2443
7	442	1877	2355	424	1860	2362
8	472	1837	2401	462	1807	2381
9	505	1791	2372	476	1783	2355
10	530	1731	2375	495	1760	2391
11	566	1692	2390	513	1732	2347
12	594	1657	2392	539	1681	2253

2.2. *Subjects, procedure and analysis*

The two authors served as subjects. DW is a native speaker of American English, BR of German. Each subject listened to 9 randomized blocks of 48 stimuli (4 repetitions of the 12 stimuli along a continuum) for each of his two personal stimulus sets. Following the design of our earlier study, each stimulus was either preceded (−500 ms stimulus onset asynchrony) or followed (750 or 3000 ms) by a 100 ms, 1000 Hz-tone, with three stimulus blocks assigned to each of these three conditions in a counterbalanced order. The subjects rapidly imitated the stimulus vowel after hearing the tone if the tone followed ("delayed" and "deferred" imitation conditions) or after hearing the stimulus if the tone preceded ("immediate" imitation condition).

In a separate test conducted several months later, each subject also identified the stimuli in his own [i]–[æ] set, using the phonemic labels /i, ɪ, e, ɛ, æ/. This test consisted of 10 randomized blocks of the 12 stimuli (without accompanying tones).

The only design change from the earlier study was that, foregoing an absolute identification (numerical labeling) task [8], each subject produced a series of isolated vowels by reading from a list containing the symbols /u, i, ɪ, e, ɛ, æ/ 36 times in random order. These productions were to serve as "prototypical" reference points in interpreting the imitation data.

The recorded imitation responses were digitized at 10 kHz, low-pass filtered at 4.9 kHz, and

[1] To simplify the analysis of stimulus-response relationships, acoustic parameter values were averaged across the whole duration of both stimulus and response vowels. The stimuli were not perfectly steady-state, however, although they represented imitations of truly stationary synthetic vowels. Formant measurements obtained at two specific points in each vowel—at onset and two-thirds into its duration—provided an indication of changes over time. These changes were relatively small and showed no orderly trends across the continua. In general, the frequencies of all formants and of the fundamental frequency declined through each vowel, except for F_1 on the [i]–[æ] continua, which tended to rise. Most of the changes in F_1 were less than 20 Hz; in F_2 and F_3, less than 100 Hz; in F_0, less than 10 Hz. Only a few tokens exceeded these limits. Clearly, none of the stimulus vowels resembled diphthongs.

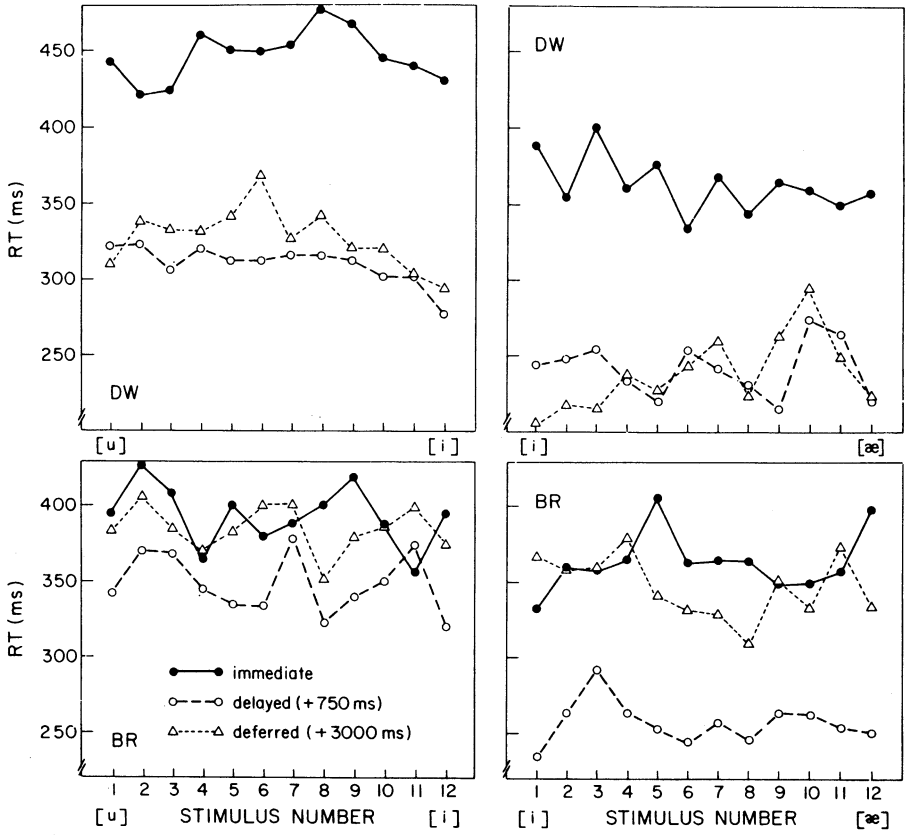

Fig. 1. Average response latencies as a function of stimulus number and delay condition for two subjects (DW, BR) and two continua ([u]–[i], [i]–[æ]). Each data point represents 12 responses.

subjected to LPC analysis.[2] The formant frequency estimates were edited to eliminate spurious and missing values, and were averaged across the whole duration of each response vowel. Mean formant frequencies and standard deviations across repeated imitations of the same stimulus were determined, as well as the distributions of formant frequencies across all responses to a given continuum. The prototypical productions were analyzed similarly. Imitation response latencies were also measured and will be discussed first.

3. Results and Discussion

3.1. Latencies

Chistovich et al. [1] observed that imitation latencies, unlike the latencies of phonetic labelling responses, did not vary systematically across an acoustic vowel continuum, regardless of response

[2] The peak-picking algorithm used to estimate formant frequencies (part of the ILS package, Version 4.0, distributed by Signal Technology, Inc.) may produce artificial discontinuities when tracking formants in time-varying signals, due to rounding errors in the FFT routine. To make sure that the present, relatively steady-state vowels had been correctly analyzed, the data from DW's [u]–[i] condition were re-analyzed using the root-solving method included in ILS, which is more accurate but time-consuming. The results were practically identical to those obtained with the peak-picking method, except that F_1 estimates were uniformly higher by about 10 Hz. The reason for this absolute difference is not known. The peak-picking algorithm thus seems to provide accurate results for relatively steady-state speech sounds.

delay. Relative uncertainty about phonemic category membership thus did not seem to influence the speed of imitation. This finding, which suggests that imitation is not mediated by phonemic classification, was essentially replicated in our earlier study [8]. The average response latencies from the present experiment are shown in Fig. 1 as a function of subject (top vs. bottom panels), continuum (left vs. right panels), stimulus number (abscissa), and delay condition (three functions). Two findings are apparent. First, although reaction times varied somewhat across each continuum, there was no consistent pattern to this variation. In other words, there were no peaks in the latency functions associated with phonetic category boundaries. Second, subject DW showed markedly slower reaction times in the immediate imitation condition than in the delayed or deferred imitation conditions, whereas subject BR showed slower latencies in the immediate and deferred conditions than in the delayed condition. While slower reaction times in the immediate imitation condition are expected because of the subjects' incomplete articulatory preparation, only BR was affected by a 3-second response delay. This pattern of results is remarkably similar to that obtained in our earlier study with synthetic stimuli.[3]

Separate repeated-measures analyses of variance with the factors Stimulus Number and Delay Condition were conducted on the average latencies for the three stimulus blocks of each continuum and each subject. The only significant effect involving Stimulus Number was a small main effect for DW on the [u]–[i] continuum [$F(11, 72) = 2.11$, $p = 0.0304$], which was not readily interpretable; the other three main effects and the four interactions were nonsignificant, which suggests the absence of reliable peaks in the latency functions. The main effect of Delay Condition, however, was highly significant ($p < 0.0001$) in each of the four analyses.

[3] Only the absolute reaction times differed: Relative to the reaction times to synthetic stimuli, DW speeded up on the [i]–[æ] continuum while BR slowed down on the [u]–[i] continuum. These changes are difficult to interpret and are of little theoretical interest.

3.2. Formant frequencies

As in our earlier study, we found that the patterns of average response formant frequencies were extremely similar across the three delay conditions, so the data were collapsed across delays.[4] The mean values were thus based on 36 responses per stimulus. These means are plotted as a function of stimulus number in Fig. 2 (solid lines); the dashed lines connect the stimulus formant frequencies.[5]

Compared to our earlier results with synthetic stimuli, the response formant frequencies are much closer to those of the stimuli, as should be expected when subjects imitate their own vowels. Nevertheless, there appear to be systematic deviations that echo some of the response nonlinearities observed with synthetic stimuli. Many of these deviations are significant individually, since standard errors are small (one-sixth of the standard deviations displayed in Fig. 5). They are also significant overall, as is clear from the results

[4] To justify this decision, analyses of variance were conducted on stimulus block mean values of F_1, F_2, and F_3 for each subject and continuum, with the factors Stimulus Number and Delay. A significant interaction between these factors would indicate a change of formant pattern as a function of delay condition. Of the twelve interactions tested, only one was significant, for F_1 along the [u]–[i] continuum of subject BR [$F(22, 72) = 1.99$, $p = 0.0157$], which is of little interest because responses to that continuum were analyzed primarily in F_2-F_3 space. The main effect of Delay was significant in several instances, indicating changes in absolute formant frequencies across delays without a concomitant change in stimulus-response relationships. The more striking of these included lower F_1 frequencies (subject DW) and lower F_2 frequencies (subject BR) in the immediate imitation of stimuli from the [i]–[æ] continuum.

[5] The responses, like the stimuli, were examined for changes in formant frequencies and F_0 over time by comparing measurements taken at vowel onset and after two-thirds of its duration. This analysis revealed that the response vowels were monophthongal and, in fact, rather stationary. The mean response parameters exhibited a variety of systematic trends in within-vowel changes across each continuum, but the magnitudes of these changes were rather small (generally less than 25 Hz for F_1, 65 Hz for F_2, 40 Hz for F_3, 16 Hz for F_0). Of the 16 stimulus-response correlations of frequency changes (4 parameters, 2 continua, 2 subjects) 15 were positive, but only one was significant. Thus there was no strong evidence that the subjects imitated time-varying characteristics of the stimuli.

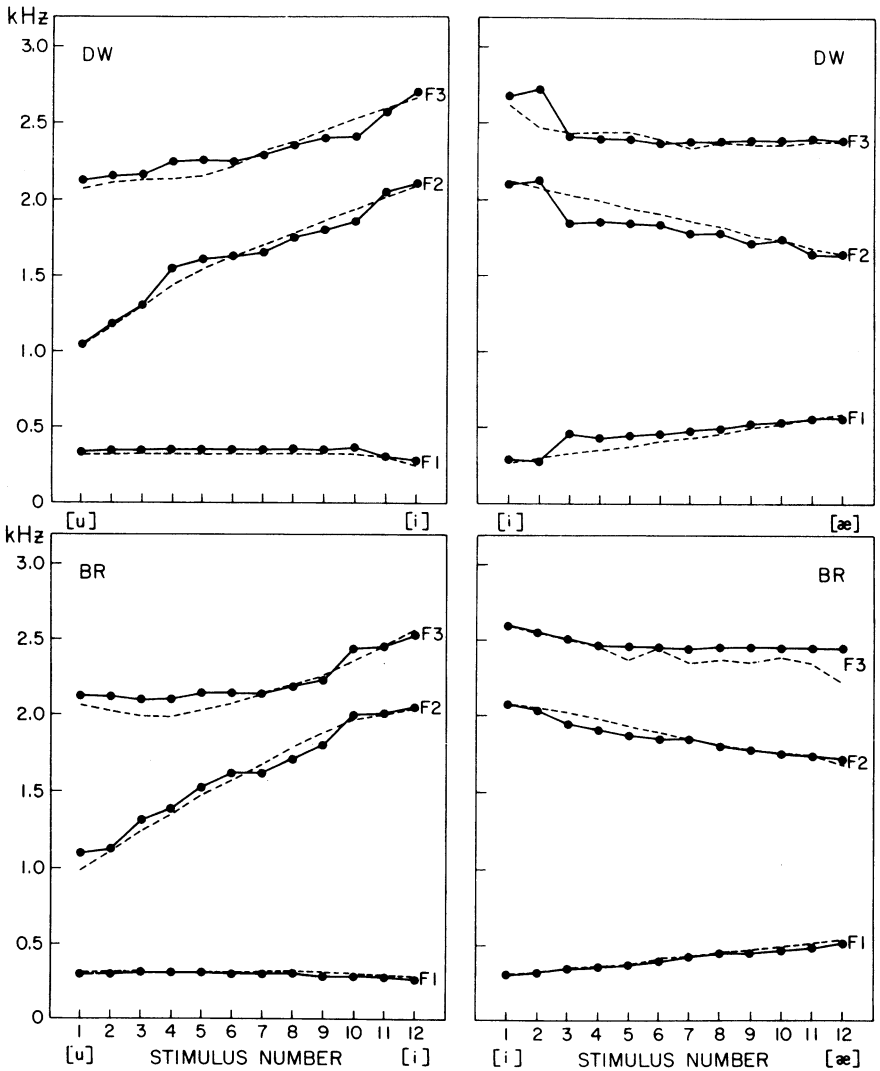

Fig. 2. Average formant frequencies of the responses as a function of stimulus number (filled circles, solid lines). Each data point represents 36 responses. The stimulus formant frequencies are connected by the dashed lines.

of analyses of variance on the deviations of the responses from the stimulus parameters. Four such analyses were conducted (two continua for each of two subjects) on three parameters (F_1, F_2, F_3) considered jointly (using a multivariate statistic) and separately. Of the grand mean effects, which test the average stimulus-response difference on each continuum, all 4 multivariate and 11 of the 12 univariate F values were highly significant ($p < 0.0001$; exception: F_2 for DW on the [u]–[i] continuum, which was nonsignificant). More importantly, all 16 stimulus number main effects, which test whether responses deviated *nonuniformly* from the stimuli across each continuum, were highly significant ($p < 0.0001$). Thus there is ample statistical support for stimulus-response nonlinearities in the data. These nonlinearities are examined more closely in the next two figures.

Figure 3 shows stimulus-response relations in

F_2-F_3 space for the [u]–[i] continuum. DW's responses to stimuli 4–10 on this series tend to cluster together, though he was able to imitate their distinctive characteristics to some extent. A similar, but weaker tendency is exhibited by BR for stimuli 5–9; in addition, BR tended to respond categorically to the endpoint stimuli (1, 2 and 10, 11, 12, respectively). These tendencies are similar to those observed in our earlier study.

Figure 4 shows the stimulus-response mapping in F_1-F_2 space for the [i]–[æ] continuum. DW shows very dramatic deviations here. There is a huge gap between the responses to stimuli 2 and 3, and responses to stimuli 3–9 are transposed down along the F_1-F_2 regression line. (Note that the responses, like the stimuli, continue to observe this linear relationship despite the large discrepancies.) There is also evidence for some endpoint clustering (stimuli 1, 2, and 11, 12, respectively). Subject BR, by contrast, shows relatively continuous responses to this continuum, although there is some contraction of the response space for stimuli 3–12. Once again, these patterns show similarities to those we have observed with synthetic stimuli. The similarities are difficult to quantify, however, because the stimuli in the two studies are not in one-to-one correspondence.

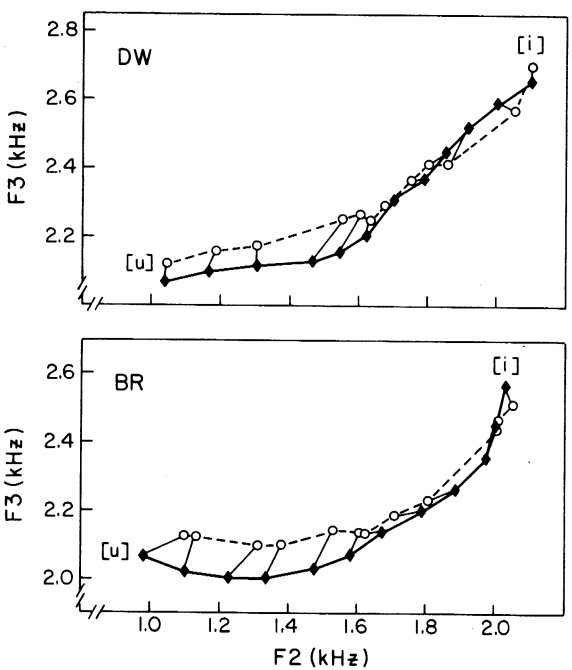

Fig. 3. Average formant frequencies of responses to the [u]–[i] continuum in F_2-F_3 space (open circles, dashed line). Filled diamonds connected by a solid line represent the stimuli. Each stimulus is connected to its corresponding average response.

3.3. Standard deviations

Another way to look for categorical tendencies is to examine the patterns of response variability. Response variability is expected to increase at category boundaries, if there are any. Standard deviations of formant frequencies, computed within but averaged across delay conditions, are shown

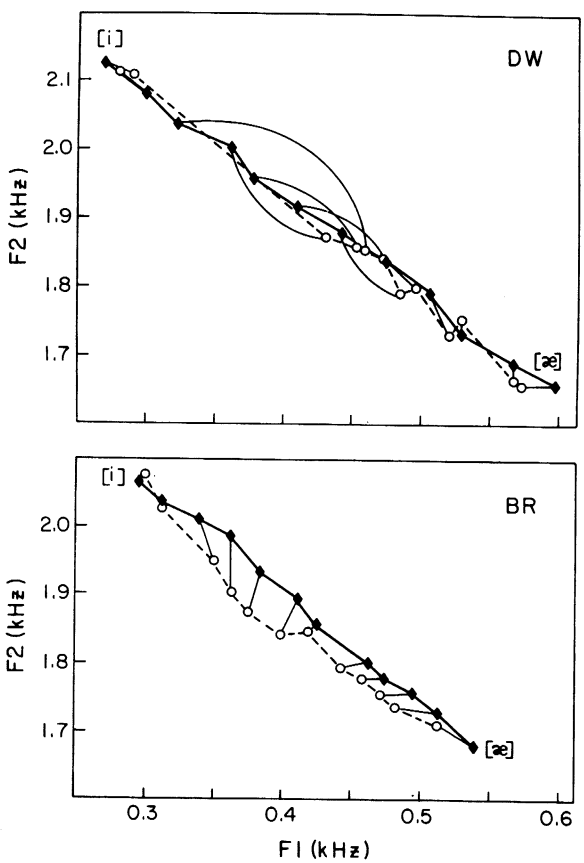

Fig. 4. Average formant frequencies of responses to the [i]–[æ] continuum in F_1-F_2 space (open circles, dashed line). Filled diamonds connected by a solid line represent the stimuli. Each stimulus is connected to its corresponding average response; the curving connectors in the upper panel are necessitated by the large response shifts.

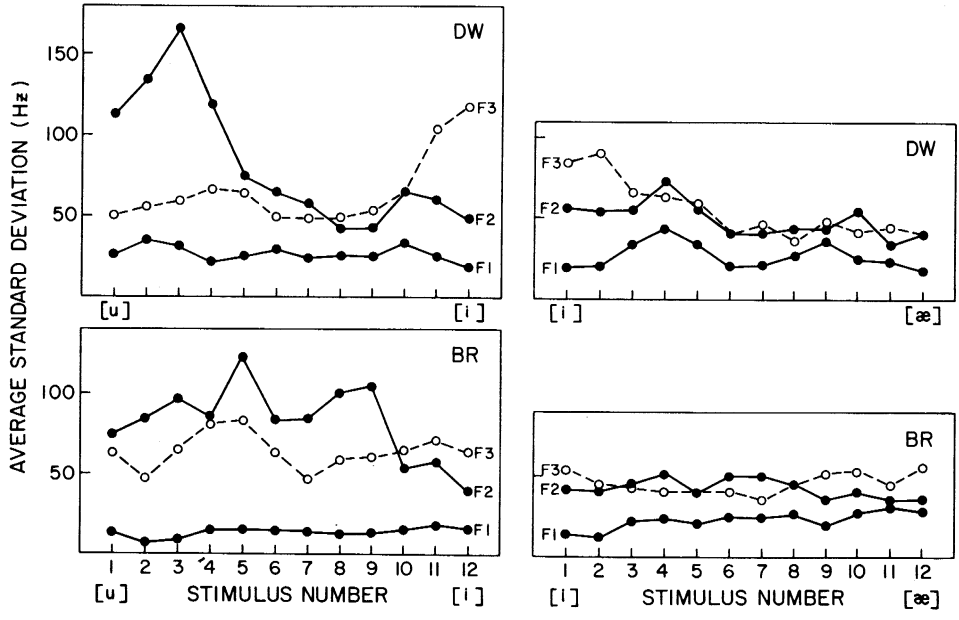

Fig. 5. Average standard deviations of response formant frequencies.

in Fig. 5. These patterns are remarkably similar to those observed with synthetic stimuli. Both subjects showed higher F_2 variability along the [u]–[i] than along the [i]–[æ] continuum, except at the [i] end. For BR, F_2 variability was elevated across most of the [u]–[i] continuum (stimuli 1–9), whereas DW showed elevated variability over a narrower region (stimuli 1–4), with a pronounced peak for stimulus 3. This peak corresponds to the gap in the formant frequency plot (Fig. 3). Apart from this feature, there are no clear indications of a categorical structure in the standard deviations along the [u]–[i] continuum. Along the [i]–[æ] continuum, however, subject DW shows two peaks in both the F_1 and F_2 functions, which suggest a three-category structure. As in the earlier study, F_1 and F_2 standard deviations were correlated for DW ($r = 0.55$, $p < 0.05$) but not for BR ($r = 0.03$). For BR, therefore, the standard deviations do not reveal any obvious categorical tendencies. Individual differences aside, however, the point to be stressed is that the standard deviations follow the same pattern as in the earlier study, suggesting that the subjects responded similarly to synthetic and natural stimuli.

3.4. Formant frequency distributions

The best way to assess categorical response tendencies is to plot overall formant frequency distributions. Frequency histogram envelopes of the first three formants of the responses in all three delay conditions combined ($n = 432$ in each graph) are shown in Figs. 6 and 7 (solid lines). For comparison, the histogram envelopes from our earlier study with synthetic stimuli are plotted alongside on the same scale (dashed lines). Significant similarities are evident.

On the [u]–[i] continuum (Fig. 6), the only major discrepancy between the two sets of results is the presence of a second peak in DW's F_1 distribution for natural speech stimuli. The cause for these unusually high F_1 frequencies in many of DW's responses is unknown. (Stimulus F_1 frequencies ranged from 276 to 313 Hz; see Table 1). BR has a single-peaked F_1 function whose displacement with respect to the earlier study brings it in good agreement with the stimulus range and corrects a consistent F_1 "overshoot" observed with synthetic stimuli. The F_2 frequency distributions of both subjects are rather similar to those obtained with synthetic stimuli and show three major peaks, two probably representing the end-

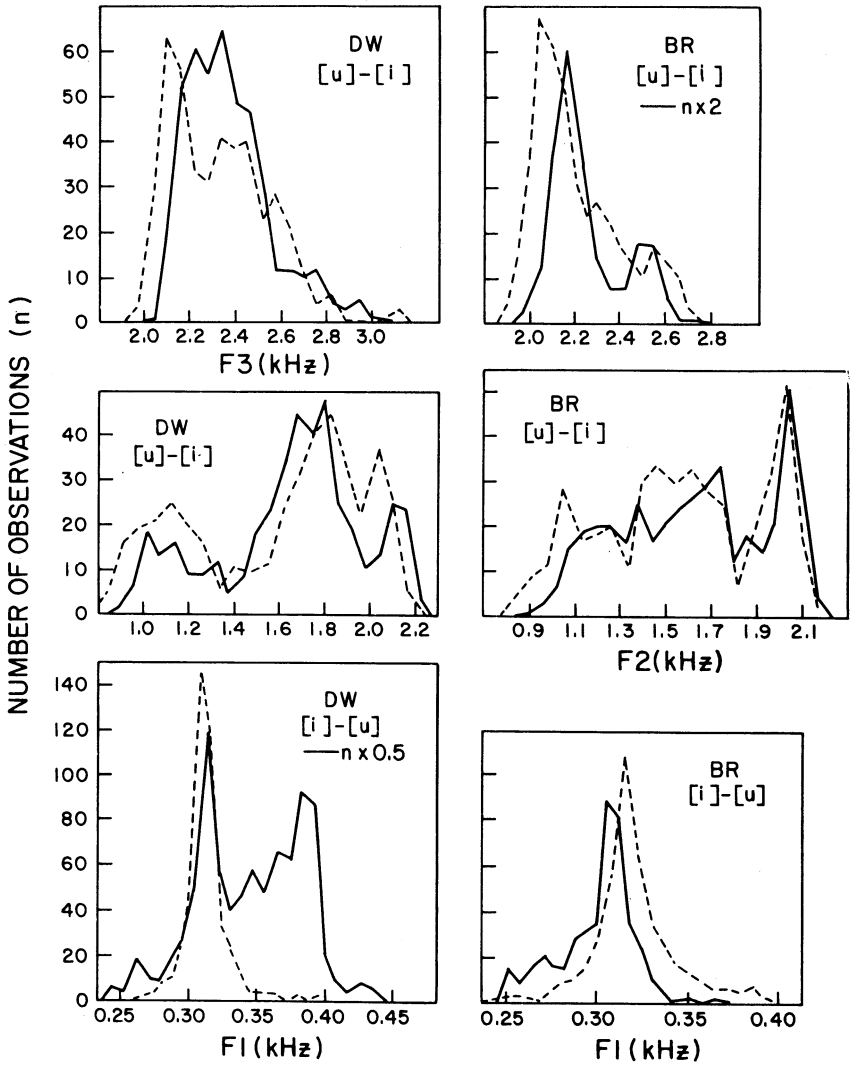

Fig. 6. Histogram envelopes of response formant frequencies for the [u]–[i] continuum in the present study (solid lines) and in our earlier study using synthetic stimuli (dashed lines). Note that the plots for the three formants are not aligned with each other, and that the scale factor is altered for some individual functions to make the functions similar in height.

point categories and the third a broad category of "unfamiliar" vowel sounds. The F_3 distributions are essentially unimodal and shifted to the right with respect to the previous study, resulting in a better match of stimulus and response F_3 ranges (cf. Table 1).

For the [i]–[æ] continuum (Fig. 7), both F_1 and F_2 show highly irregular distributions indicative of categorical tendencies, whereas the F_3 distribution is unimodal. For DW, both the F_1 and F_2 distributions are trimodal; moreover, the peaks (taking into account the reversal of the continuum along the F_2 scale) are in fact aligned with each other. DW thus shows evidence for three categories along this continuum. For BR, the pattern is less clear. The F_1 histogram shows four peaks, adding a new one to the three-peaked function for synthetic stimuli. The F_2 function has multiple peaks—too many for any clear categorical structure to be inferred.

The main result of these comparisons is that individual response preferences are maintained to

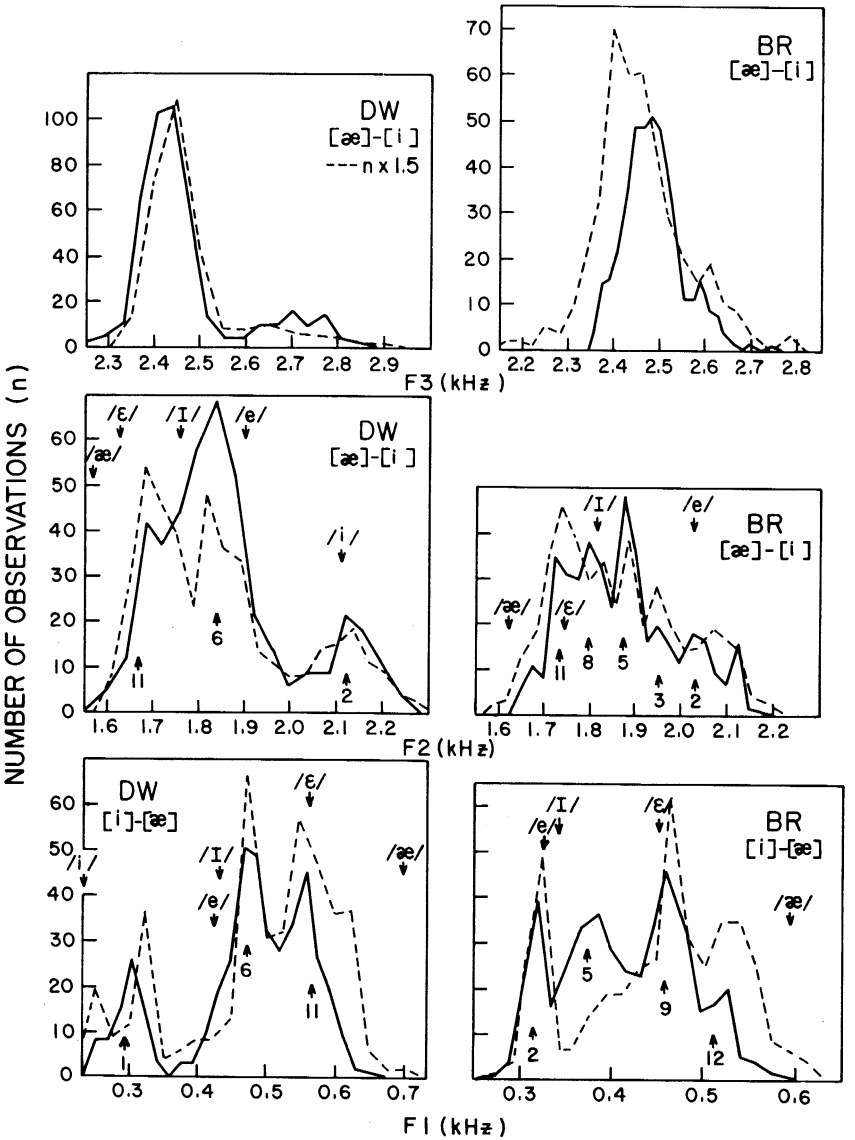

Fig. 7. Histogram envelopes of response formant frequencies for the [i]–[æ] continuum in the present study (solid lines) and in our earlier study using synthetic stimuli (dashed lines). Note that the plots for the three formants are not aligned with each other, that the continuum is reversed for F_2 and F_3 with respect to F_1, and that the scale factor is altered for the dashed function in the upper left-hand panel. Arrows with numbers represent the stimuli whose average responses fell closest to histogram peaks. Arrows with phonetic symbols represent prototypical vowel productions.

a considerable extent even when subjects imitate self-produced vowels. Clearly, few of the distributions are uniform, as they should be if formant frequencies were reproduced faithfully.

3.5. *Phonemic identification*

The subjects labeled the stimuli along their own [i]–[æ] continua to provide a reference for the interpretation of categorical tendencies along that continuum. These classifications are plotted

in Fig. 8. It can be seen that DW used only three categories (/i, e, ɛ/) consistently; he used /æ/ interchangeably with /ɛ/, and /ɪ/ not at all. That is, for him the stimulus continuum represented only three categories. BR, on the other hand, applied all five response categories to his vowels, although stimulus 12 still was only a weak /æ/ to him.

To see whether these data are helpful in interpreting the histogram peaks, the ordinal numbers of the stimuli whose associated mean response formant frequencies were close to histogram peaks have been entered below arrows in Fig. 7. For subject DW, the three major peaks in F_1 and F_2 are associated with responses to stimuli identified as /i/, /e/, and /ɛ/ (or /æ/), respectively. This correspondence is in agreement with that observed in our earlier study, except that we then interpreted the /e/ category as /ɪ/. DW's categorical tendencies in imitation thus correspond well to his phonemic categories. For BR, the F_1 and F_2 peaks line up with stimuli labeled as /i/, /ɪ/, /ɛ/, and /æ/, respectively, although there seem to be two /i/ peaks in the F_2 distribution. These alignments differ somewhat from those obtained in our earlier study and therefore must be regarded with caution. That BR, as a native speaker of German, should not have a well-defined /e/ category in imitation seems counterintuitive. For this subject, then, the imitation data are not clearly related to his (English) phonemic categories, perhaps because of his bilingualism.[6]

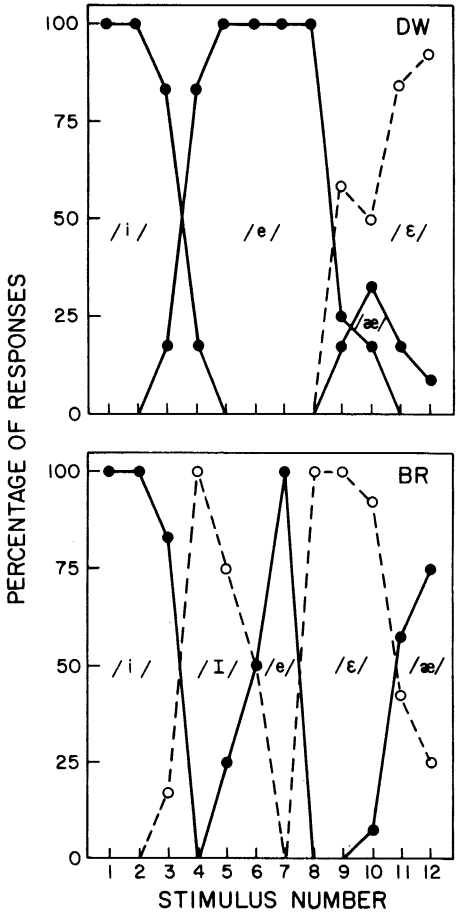

Fig. 8. Labeling responses to the [i]–[æ] continuum.

3.6. *Prototypical vowels*

A new feature of the present study was the inclusion of "prototypical" productions representing the five English vowel categories along the [i]–[æ] continuum. The average frequencies of these productions have been entered above arrows in the F_1 and F_2 panels of Fig. 7. Somewhat surprisingly, these values are not very helpful in interpreting the histogram peaks. The prototypical values for /æ/ generally fall outside the response ranges. Those for the other categories generally do not coincide with major peaks, although some tentative alignments can be made if small shifts in formant frequencies are allowed for. Clearly, the subjects did not simply produce their prototype vowels in the imitation task. Their responses definitely were more a function of the stimuli than of pre-established phonetic categories, although the categories may have exerted a certain "pull" on the responses.

To get a better idea of the locations of the prototype vowels in the formant frequency plane relative to the stimulus and response vowels, the

[6] Because of these puzzling results, we later repeated the identification task, also with the two subjects listening to each other's [i]–[æ] series. This replication revealed considerable inconsistency in the subjects' use of the /ɪ/ category, and both subjects agreed that no very good instances of this vowel were present in either stimulus series. BR's data make more sense if stimulus 5, and the associated peaks in the F_1 and F_2 histograms, are taken to represent his /e/ category.

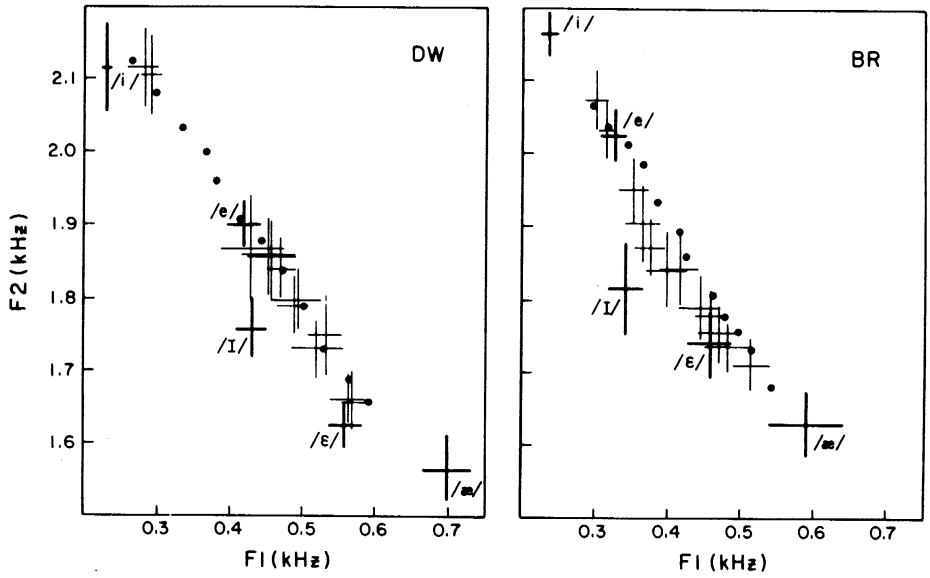

Fig. 9. Average formant frequencies of responses along the [i]–[æ] continuum plus/minus one standard deviation (thin lines) and of prototypical vowel productions plus/minus one standard deviation (heavy lines) in F_1-F_2 space. The circles represent the stimuli.

subjects' responses to their [i]–[æ] continua have been replotted in Fig. 9 together with the prototypes, with standard deviations represented as well. The stimuli appear as filled circles. One interesting feature emerging from these plots is that, for both talkers, the five prototype vowels do not lie on a straight line in F_1-F_2 space, in contrast to the response (and stimulus) vowels. It seems that the subjects, being rather accurate imitators, fit their responses to the linear trajectory imposed by the stimuli, rather than gravitating toward their prototypical vowels. Prototypical /ɪ/, in particular, lies outside the stimulus-response trajectory, and /æ/, as well as BR's /i/, is beyond the stimulus-response range. Most responses fall between prototypical /e/ and /ɛ/; only DW also produced some /i/-like vowels. The main difference between the two subjects is in the location of the /e/ prototype, which is closer to /i/ for BR and presumably reflects his native language. The absence of a prototype for DW in the same region may explain the large shifts in his responses to stimuli 3–5. Curiously, BR labeled stimuli as /i/ (Fig. 8) that in fact were much closer to his prototypical /e/, and the stimuli he labeled as /e/ were closer to his prototypical /ɪ/. DW's labeling responses are in much better agreement with the pattern of stimulus-prototype proximities shown in Fig. 9[7].

3.7. *Fundamental frequencies*

We examined two additional stimulus-response relationships that we could not explore in our earlier study because of the constant fundamental frequency (F_0) and duration of the synthetic stimuli. First, we compared the average fundamental frequencies of the stimuli and of the responses. For DW, there were no major trends in response F_0 across either continuum; occasional

[7] See, however, footnote 6. BR's labeling data from the replication were in somewhat better agreement with his prototypes. Also, both subjects' productions of /ɪ/ may have been anomalous; after all, this English vowel does not occur in isolation. As a matter of fact, both subjects' productions of all vowels deviate considerably from the Peterson-Barney norms [5], which are based on vowels produced in /h_d/ context (not including /e/). It should also be mentioned that DW's prototypical productions, but not BR's, tended to be diphthongized. Both subjects' ability to identify their own and each other's prototypes was tested later. Scores ranged from 93 to 100 percent correct, with most confusions involving intended /ɪ/ or /ɛ/.

deviations seemed to be related to stimulus F_0. Stimulus-response correlations in the three delay conditions for each continuum ranged from 0.39 to 0.88 (4 out of 6 significant at $p < 0.01$), which indicated that DW unintentionally imitated stimulus F_0. For BR, the correlations were lower but still positive, ranging from 0.25 to 0.62 (1 out of 6 significant at $p < 0.05$), and his response F_0 tended to fall across both continua (from [u] to [i], and from [i] to [æ]), an effect that was apparently not induced by the stimuli. Stimulus-response correlations for both subjects tended to be lower in the immediate imitation condition. Delay conditions affected absolute F_0, but these patterns varied between subjects and continua and were difficult to interpret.

3.8. Durations

Similarly, we examined stimulus and response durations along each continuum and found some very consistent patterns. The stimulus-response correlations were positive and surprisingly high in some instances. For DW, they ranged from 0.30 to 0.96 (5 out of 6 significant at $p < 0.01$); for BR, from 0.35 to 0.70 (4 out of 6 significant at $p < 0.05$, one of those at $p < 0.01$). Although it might be argued that a common articulatory or phonetic factor influenced stimulus and response durations alike, the pattern of durations across each continuum was sufficiently irregular (due to the method of stimulus selection) to suggest, rather, that both subjects unintentionally mimicked vowel durations. The stimulus-response correlations tended to be lower in the deferred imitation condition. In addition, there was a very pronounced effect of delay condition on the average duration of the responses: Response vowels were generally shorter in the immediate imitation condition.

4. Conclusions

On the whole, the present results replicate the findings of our first study [8]. That is, categorical tendencies in vowel imitation are obtained even when the subjects are capable of producing the precise vowel they are to imitate. This rules out one possible explanation of the obtained stimulus-response nonlinearities, namely, that they arise in the translation of nonproduceable stimuli into the subject's own production space. As pointed out in the Introduction, this hypothesis had limited plausibility to begin with; thus a sample of two subjects seems sufficient for its dismissal. At the same time, the demonstration of similar nonlinearities with synthetic and natural stimuli confirms the robustness of these effects, as well as the presence of considerable individual differences in their pattern and magnitude (cf. also [4]).

One possible reason for the absence of very strong categorical effects in this study and its predecessors [4, 8] is suggested by the relation of the subjects' prototypical vowels to the [i]–[æ] stimulus continuum. Our continuum derived from responses to a synthetic continuum [8], which we had copied from Kent [4], who in turn had designed it to span the average male vowel formant frequencies for /i/ and /æ/ reported by Peterson and Barney [5]. These latter data derived from vowels in /h__d/ context and may not be representative of isolated vowel productions (especially /ɪ/ and /ɛ/), for which normative English data are hard to come by in the literature. It is also possible that the present subjects were not representative of the average American male talker. In any case, it seems that the [i]–[æ] continua used by Kent and by us did not span the full space between /i/ and /æ/, and that they bypassed /ɪ/. Chistovich et al. [1] used a continuum that seems to have been more closely matched to their single subject's prototypes, and it remains to be seen whether their highly categorical results can be replicated with similarly constructed stimulus continua.

The question of the origin of categorical tendencies in vowel imitation needs to be addressed in further research. Perhaps the most interesting result to emerge from our studies and that of Chistovich et al. [1] is that categorical tendencies in imitation appear regardless of response delay (up to 2 seconds) and with essentially constant reaction times. Imitation responses thus do not seem to be mediated by explicit phonemic decisions (which are slowed by stimulus ambiguity), nor do they depend on a rapidly decaying auditory memory (which plays a role in vowel dis-

crimination, see [2, 3, 6]). This suggests that the internal representation of perceived vowels is phonetic (or articulatory) but, at the same time, either noncategorical or only weakly categorical. If it is noncategorical, then the categorical tendencies must arise during the motor implementation of the imitations. Research is now in progress to examine this possibility.

Acknowledgments

This research was supported by NICHD Grant HD-01994 and BRS Grant RR-05596 to Haskins Laboratories. Portions of the results were reported at the 109th meeting of the Acoustical Society of America in Austin, TX, April 1985.

References

[1] L.A. Chistovich, G. Fant, A. de Serpa-Leitão, and P. Tjernlund, "Mimicking and perception of synthetic vowels", *Quarterly Progress and Status Report* (Royal Technical University, Speech Transmission Laboratory, Stockholm), No. 2, 1966, pp. 1–18.

[2] R.G. Crowder, "Decay of auditory information in vowel discrimination", *J. Experimental Psychology: Human Learning, Memory, and Cognition*, Vol. 8, 1982, pp. 153–162.

[3] R.G. Crowder, "A common basis for auditory sensory storage in perception and immediate memory", *Perception and Psychophysics*, Vol. 31, 1982, pp. 477–483.

[4] R.D. Kent, "The imitation of synthetic vowels and some implications for speech memory", *Phonetica*, Vol. 28, 1973, pp. 1–25.

[5] G.E. Peterson and H.L. Barney, "Control methods used in a study of the vowels", *J. Acoust. Soc. Am.*, Vol. 24, 1952, pp. 175–184.

[6] D.B. Pisoni, "Auditory short-term memory and vowel perception", *Memory and Cognition*, Vol. 3, 1975, pp. 7–18.

[7] B.H. Repp, A.F. Healy and R.G. Crowder, "Categories and context in the perception of isolated, steady-state vowels", *J. Experimental Psychology: Human Perception and Performance*, Vol. 5, 1979, pp. 129–145.

[8] B.H. Repp and D.R. Williams, "Categorical trends in vowel imitation: Preliminary observations from a replication experiment", *Speech Communication*, Vol. 4, 1985, pp. 105–120.

[9] K.N. Stevens, "The quantal nature of speech: Evidence from articulatory-acoustic data", in: E.E. David and P.B. Denes, eds., *Human Communication: A unified view*, McGraw-Hill, New York, 1972, pp. 51–66.