Speech Communication 4 (1985) 105–120
North-Holland

CATEGORICAL TRENDS IN VOWEL IMITATION: PRELIMINARY OBSERVATIONS FROM A REPLICATION EXPERIMENT

Bruno H. REPP and David R. WILLIAMS*

Haskins Laboratories, 270 Crown Street, New Haven, CT 06511-6695, USA

Received 5 November 1984

Abstract. The question whether isolated stationary vowels are imitated in a continuous or categorical fashion—raised by Chistovich et al. [2] and followed up primarily by Kent [4]—was pursued further by replicating Kent's study with some slight modifications. The subjects (the two authors) imitated synthetic stimuli from [u]-[i] and [i]-[æ] continua at three different temporal delays. Acoustic analysis of the response vowels revealed very similar patterns across delays. Both subjects showed clear evidence of nonlinearities in the stimulus-response mapping and of preferred response formant frequencies, though strictly categorical responses were generally absent. The origin of these nonlinearities and their relation to the phonemic vowel categories of English are not fully understood at present. Vowel imitation responses presumably reflect the joint influences of perceptual and articulatory factors that need to be disentangled in future research.

Zusammenfassung. Die von Chistovich et al. [2] aufgeworfene und später von Kent [4] behandelte Frage, ob isolierte gehaltene Vokale in kontinuierlicher oder kategorischer Weise imitiert werden, wurde in an Kents Ansatz angelehnten Untersuchungen neu aufgegriffen. Die Versuchspersonen (beide Autoren) imitierten synthetische Stimuli zwischen [u]-[i] und [i]-[æ] bei drei unterschiedlichen zeitlichen Verzögerungen. Die durch die akustische Analyse der imitierten Vokale erhaltenen Muster erwiesen sich unter allen Verzögerungsbedingungen als ausgesprochen ähnlich. Für beide Versuchspersonen zeigten sich deutliche nicht-lineare Beziehungen zwischen den Stimuli und den entsprechenden Reaktionen sowie bevorzugte Formantfrequenzen, obgleich im strengen Sinne kategorische Reaktionen nicht nachzuweisen waren. Eine Erklärung dieser Nicht-Linearitäten und ihrer Beziehung zu phonemischen Vokal-Kategorien des Englischen erscheint zum jetzigen Zeitpunkt noch nicht möglich. Verbale Reaktionen bei Vokalimitationsaufgaben spiegeln wahrscheinlich gemeinsame Einflüsse perzeptueller und artikulatorischer Faktoren wider, deren Trennung zukünftigen Untersuchungen vorbehalten bleibt.

Résumé. La question de savoir si des voyelles stationnaires isolées sont imitées de manière continue ou catégorielle a été soulevée par Chistovich et al. [2] et réexaminée par la suite par Kent [4]. Nous avons examiné ce problème en rédupliquant l'étude de Kent avec de légères modifications. Les sujets (les 2 auteurs) ont imité des stimuli synthétiques variant entre [u]-[i] et [i]-[æ] à trois détails temporels différents. L'analyse acoustique des réponses montre que les configurations obtenues pour les trois délais sont trés similaires. Pour les deux sujets, des non-linéarités apparaissent clairement dans les relations stimulus-réponse et dans les fréquences formatiques préférentielles, bien que des réponses strictement catégorielles soient généralement absentes. L'origine de ces non-linéarités et leurs relations avec les catégories vocaliques de l'anglais ne sont pas entièrement comprises jusqu'à présent. Les réponses d'imitation vocalique reflètent vraisemblablement des influences conjuguées de facteurs perceptifs et articulatoires qui doivent être déconvolués dans des recherches futures.

Keywords. Speech imitation, vowel production, categorical perception.

* Also Department of Psychology, University of Connecticut, Storrs, CT 06268.

1. Introduction

Almost two decades ago, Chistovich and her colleagues [2] raised the interesting possibility that (isolated, stationary) vowels might have an internal representation intermediate between the auditory pattern, which presumably is a continuous function of the input, and the discrete categories of the vowel phonemes in the language. To warrant a separate existence in a theoretical model of speech processing, such an intermediate level of representation must have a noncontinuous structure distinct from that found at the phonemic level.

To test for this possible intermediate level, Chistovich et al. [2] used a vowel imitation task. There is reason to believe that oral reproduction, particularly when it occurs as rapidly as possible ("shadowing"), may bypass the level at which familiar phonemic categories are associated with the input. This may be so even when the imitation response occurs at some delay ("mimicking"):[1] Latencies of shadowing and mimicking responses do not increase for phonemically ambiguous vowel tokens, but the latencies of written responses do [2]. The increase presumably reflects the time needed to decide among several categorical response alternatives. The hypothetical intermediate stage thus may serve to translate auditory information directly into motor instructions, as well as to store such information over time periods exceeding the life span of the raw auditory trace. Indeed, this intermediate representation may be thought of as motor in nature, thus anchoring it firmly in one of the three essential components of any model of speech communication (sensory, motor, and central processing). The question posed by Chistovich et al. thus concerns the continuous versus discrete motor representation of (isolated, stationary) vowels.

To answer this question, Chistovich et al. [2] constructed a series of 12 synthetic vowel stimuli forming a continuum between the Russian vowel categories [i]-[ɛ]-[a]. The formant values of the endpoint stimuli were closely matched to natural productions of the single subject (L.C.), a female speaker of Russian. Her imitation responses in two conditions (shadowing—average latency of 190 ms; mimicking—average latency of 900 ms) were analyzed in three ways: (1) The formant frequencies (*F*1, *F*2, *F*3) of the reponses, obtained from spectrograms, were plotted as a function of the formant frequencies of the synthetic stimuli. (2) The standard deviations of the formant frequencies across multiple imitations of the same stimulus were plotted in a similar fashion. (3) Histograms of the frequencies of each formant across all responses were constructed.

For the mimicking task, the formant frequency plots revealed stepwise changes in *F*2 suggesting at least four distinct categories. This impression was supported by the presence of three peaks in the standard deviation plot, corresponding to the boundaries between those categories. Finally, the *F*2 histograms also seemed to represent four distinct distributions of frequency values. The shadowing data seemed to agree, although only the first type of analysis was presented and some additional arguments were required to establish the similarity. Phonetic transcriptions by the (sophisticated) subject of her own responses also suggested four categories—corresponding to /i/, /ɛ/ (or /e/), /ö/, and /a/—and possibly a fifth category (/I/). Further support for this division came from a vowel matching task that required long-term memory for a fixed standard, using subject L.C. and the same set of stimuli. Chistovich et al. concluded that there is a discrete representation of vowels in terms of categories whose number exceeds that of the relevant phonemic categories in the language, and that this representation is used both to guide articulation and to retain vowel sounds in long-term memory.

Although these results are suggestive and challenging, they are not without weaknesses: (1) There was convincing evidence for only one category (/ö/) beyond the three phonemic categories

[1] The term "mimicking", used by Chistovich et al. (1966), suggests that the subject's response must match the stimulus in every respect. This was nearly true in their study, where the stimuli were modelled after vowels produced by the subject. In discussing our own data and those of Kent (1973), we will use the more general term, "imitation", which allows for stimulus-response differences in various irrelevant properties (duration, fundamental frequency, etc.) as well as for differences in formant frequencies caused by mismatches between the vocal tract implied by the stimulus and that of the imitator.

that, *a priori*, might have been expected to play a role. That extra vowel category, moreover, is functional in Swedish, the language of the country where Chistovich et al. conducted their study. Thus, the results are not inconsistent with a two-stage model, in which the outcome of a rapid phonemic decision (prior to the stage of response selection) constrains the motor program for imitation, without any intermediate stage. (2) The conclusions are based entirely on the pattern of *F*2 frequencies. Although there appeared to be some correlated trends in *F*1 and *F*3, these data were not discussed by Chistovich et al. (3) The analysis of shadowing responses is incomplete, and not totally conclusive. (4) All the analyses are qualitative; no statistical criterion for, e.g., detecting steps in a function was specified. (5) There was only a single, highly sophisticated subject.

Some preliminary follow-up data were reported by Chistovich et al. [3]. Several subjects (including L.C.) imitated noise-excited synthetic two-formant vowels lying along arbitrary trajectories in the *F*1-*F*2 space. Again, some step-wise changes in response formant frequencies were observed, but the data show some remarkable irregularities and are suggestive at best. Some of the stimuli may have exceeded the range of formant values subjects were able to produce.

A full-scale replication, with some modifications, was attempted by Kent [4]. He used two 11-member vowel continua, ranging from [u] to [i] and [i] to [æ], respectively. The formant values of the endpoint stimuli were derived from Peterson and Barney [9]. Four English-speaking subjects who had some phonetic training imitated each vowel 10 times. There was no time pressure; subjects had up to 7 s to respond.

The [i]-[æ] continuum roughly corresponds to two-thirds of the [i]-[a] continuum used by Chistovich et al. [2]. As Kent points out, this continuum is "phonemically rich" in English, spanning as many as five categories (/i/, /I/, /e/, /ɛ/, /æ/). The [u]-[i] continuum, on the other hand, contains no familiar vowel categories between the endpoints, despite covering a much larger range of *F*2 frequencies (though with *F*1 nearly constant). The question of interest, then, was whether categorical tendencies in vowel imitations—if replicated—would be restricted to the [i]-[æ] continuum.

Kent interpreted his data as providing an affirmative answer to this question. Response *F*2 frequencies along the [u]-[i] continuum, while not a strictly linear function of stimulus *F*2, did not exhibit any clear steps for any of the four subjects. The standard deviations, however, showed pronounced peaks—a single central peak for subjects 1 and 3, and twin peaks for subjects 2 and 4. Moreover, the *F*2 histograms tended to have several modes for each subject—two for subjects 1 and 3, and three for subjects 2 and 4. Finally, comparison of these patterns with the *F*2 frequency plots shows that increased standard deviations and histogram valleys correspond to regions of increased slope on the *F*2 frequency plots. Kent was very conservative in interpreting these data, conceding only that responses were more accurate at the [u] and [i] ends of the continuum. In fact, his data offer some support for the presence of three categories in two of the four subjects. It is true, however, that these categories correspond only to regions of reduced discrimination, never to a true constancy in imitation. No labeling data were collected for these stimuli, nor was *F*3 analyzed.

The results for the [i]-[æ] continuum were presented by Kent in an incomplete and somewhat confusing manner, which makes it difficult to object to his conclusion that these data were hard to interpret. The response locations in *F*1-*F*2 space (*F*3 was not analyzed) showed some pronounced discontinuities, but they occurred in different places for different subjects and agreed only partially with the patterns of standard deviations and frequency histograms. For only one subject could histogram peaks be tentatively aligned with the response categories used in labeling the [i]-[æ] stimuli. From these results, Kent drew the "very tentative conclusion that the stimuli of the /i/-/æ/ series are represented in memory in a more classificatory fashion than are the stimuli of the /u/-/i/ series" ([4], p. 16). The emphasis on memory is explained by the lack of time constraints in Kent's task.

The available data on categorical tendencies in vowel imitation must therefore be considered as merely suggestive. Both Kent [4] in his concluding paragraph and, quite recently, Chistovich [1] acknowledge that further data on this issue are

needed. Their initial studies do not seem to have been followed up.[2] Kent went on to conduct a series of interesting vowel imitation studies, but none of these addresses directly the issue of categorical tendencies [5, 6, 7, 8]. A relevant study by Schouten [12], in which stimuli from an [i]-[æ] continuum were presented to Dutch-English bilingual subjects, yielded some evidence of categorical imitation, but the data were complex and were analyzed by quite different procedures, such as cluster analysis. Thus, the ambiguity of the earlier results remains.

We decided to continue where Kent [4] left off, and to begin with a straightforward replication of his study. In this preliminary report we present data obtained for the two authors as subjects. There were a few methodological differencies between our study and Kent's. The most important of these was that we employed three different response timing conditions, in an attempt to incorporate the "shadowing" versus "mimicking" comparison [2] into Kent's design. Subjects were required to imitate each vowel (a) immediately, (b) following a short (750 ms) delay, and (c) following a longer (3 s) delay. Other procedural changes were relatively minor: First, the stimuli were presented over a loudspeaker rather than over earphones. Second, to prevent overlap of shadowing responses with the end of the stimulus, the synthetic vowels were shortened to 150 ms (versus 250 ms [4] and 300 ms [2]). Extrapolating from perceptual findings [10], shortening of vowel stimuli should, if anything, result in a more categorical response. Finally, there were 12 (rather than Kent's 11) stimuli per vowel continuum.

2. Methods

2.1. Subjects

The two authors served as subjects in this pilot study. BR is a native speaker of Southern German who has been speaking English almost exclusively for over 15 years. He also has smatterings of French, Italian, and Swedish, but no professional phonetic training. DW is a native speaker of American English from California. Although not a fluent speaker of any foreign languages, he has studied French, German, Japanese, and Latin, and has had several years of phonetic training.

2.2. Stimuli

Five-formant vowel stimuli were generated on the software synthesizer at Haskins Laboratories. The center frequencies of the fourth and fifth formants were fixed at 3500 and 4500 Hz, respectively. All stimuli were 150 ms in duration and all formants were stationary with amplitude rise-fall times of 30 ms. Bandwidths of the five formants were up at 50, 80, 110, 150, and 200 Hz, respectively. For all stimuli, the fundamental frequency fell linearly from an initial value of 120 Hz to a final value of 105 Hz.

Two stimulus continua—one from [u] to [i] and the other from [i] to [æ]—were employed in the experiment. Frequency values for *F*1, *F*2, and *F*3 of the endpoint stimulus vowels corresponded to the mean values reported by Peterson and Barney [9] for their male talkers. Two 12-stimulus vowel series were constructed by interpolating in equal frequency steps between the endpoint values for the first three formants. The formant frequency values of all stimuli are listed in Table 1. Twelve randomized blocks of the twelve stimuli were arranged to form a stimulus sequence. The order of items in the sequence was adjusted until it was almost perfectly balanced; i.e., each stimulus was preceded once by every other stimulus, with a few exceptions. For each vowel continuum, three such stimulus sequences were recorded on tape for presentation in the imitation task.

2.3. Imitation task

Each of the 144 trials in a stimulus sequence required 6 seconds. Within that interval, the subject was presented with a 150 ms stimulus vowel that was either preceded or followed by a 1000 Hz tone, 100 ms in duration. The position of the tone in a trial depended on the condition. In the *immediate* imitation condition (A), a trial began

[2] We yet have to conduct a full survey of the relevant literature, especially that published in Russian.

Table 1
Formant frequencies of the stimuli

[u]-[i] continuum			
	F1	*F2*	*F3*
1	300	870	2240
2	297	999	2310
3	295	1128	2380
4	292	1257	2450
5	289	1386	2520
6	286	1515	2590
7	284	1644	2660
8	281	1773	2730
9	278	1902	2800
10	276	2031	2870
11	273	2161	2940
12	270	2290	3010

[i]-[æ] continuum			
	F1	*F2*	*F3*
1	270	2290	3010
2	305	2238	2955
3	341	2186	2901
4	376	2134	2846
5	412	2082	2792
6	447	2030	2737
7	483	1979	2683
8	518	1927	2628
9	554	1876	2574
10	589	1824	2519
11	625	1772	2465
12	660	1720	2410

with the tone, which served as a warning signal; its onset preceded that of the stimulus vowel by 500 ms. Subjects were asked to produce a response vowel as soon as possible after the onset of the stimulus vowel. In the *delayed* (B) and *deferred* (C) imitation conditions, a trial began with a stimulus vowel, the onset of which preceded that of the tone by 750 ms and 3000 ms, respectively. Subjects were asked in these conditions to respond immediately upon the onset of the tone, which served as a "go" signal here. All responses were initiated from a closed mouth position; that is, phonation was always preceded by a silent opening gesture.

Each of the three stimulus sequences was divided into three parts of 48 trials each, separated by pauses and assigned to the three imitation conditions as follows: (1) A, B, C; (2) B, C, A; (3) C, A, B. Before hearing the stimulus sequences, subjects listened to one 12-item block of trials from each of the imitation conditions for practice. The [u]-[i] session preceded the [i]-[æ] session for both subjects.

The stimuli were presented in a sound-insulated (IAC) booth over a Realistic loudspeaker which was placed about 20 degrees to the right of center at a distance of 3 feet from the subject. A Sennheiser MKH415T microphone was placed directly in front of the subject approximately eighteen inches from his lips. Both the stimuli and the subject's responses were recorded on a second tape recorder.

2.4. Identification tasks

In addition to the imitation responses, each subject provided written phonemic and numeric identifications of the stimuli. For the phonemic identification task, the stimulus vowels from the [i]-[æ] series were recorded twelve times in random sequence with ISIs of 3 s. The subjects used the labels /i, I, e, ɛ, æ/. This identification task followed the imitation task by several weeks. Phonemic identification of the [u]-[i] stimuli was not attempted, since both authors found it difficult to think of appropriate response categories.

Several weeks later, the subjects identified the stimuli from both continua on a numerical scale ranging from 1 to 12. To facilitate this absolute identification task, the stimuli were presented in an AXB format, with the A and B stimuli representing the continuum endpoints, in either order. A total of $12 \times 12 = 144$ trials were presented for each continuum, with ISIs of 1 s within triads and 3 s between triads.

Each identification task was preceded by a familiarization sequence in which the stimulus series was presented four times, twice in forward order and twice in reverse order.

2.5. Analysis

The imitation recordings were digitized at a sampling rate of 10 kHz. With a waveform editor, the time interval between the onset of the warning tone and the onset of the response vowel was

measured. Each response vowel was then isolated, labeled, and stored in a disk file. The center frequencies of the first three spectral peaks were obtained using LPC analysis, with a 20-ms window moving in 10-ms steps. For each formant, three separate estimates were obtained: (a) the mean formant frequency over the whole vowel, including its standard deviation across the vowel token; (b) its value at the onset of the response vowel; (c) its value two-thirds into the vowel (the measurement point used in both [2] and [4]). These values were further averaged across the 12 responses to each stimulus token, and the standard deviation for each stimulus token was computed. Formant frequency histograms for each delay condition and for all delay conditions combined were obtained for each stimulus in a series using Program 5D of the BMDP package.

3. Results and discussion

3.1. Latencies

Chistovich et al. [2] reported essentially constant response latencies for subject L.C. across their vowel continuum, both in the rapid shadowing condition (average latency: 190 ms) and in the slower mimicking condition (average latency: 900 ms). Kent [4] did not measure reaction times. The average latencies for the vowels along the two continua used in the present study are plotted in Fig. 1 as a function of stimulus number, sepa-

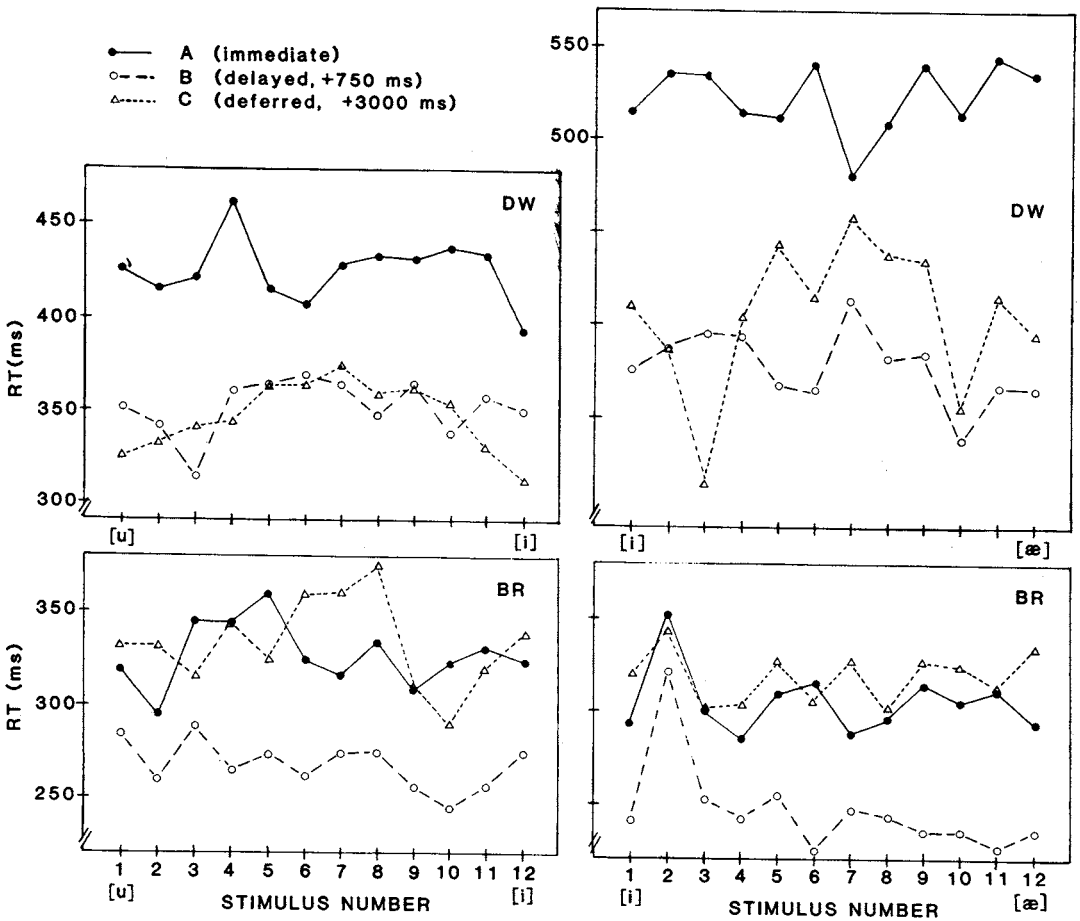

Fig. 1. Response latencies (averaged over 12 responses) as a function of stimulus number, imitation condition, vowel continuum, and subject.

rately for the two subjects and for the three imitation conditions. Although there is more variability here than in the data of Chistovich et al., the functions are either flat or vary in a nonsystematic fashion. The standard deviations were rather large—about 60 ms on the average—which must be taken into account when interpreting the latency functions. Despite occasional peaks and valleys, these functions do not provide any strong evidence of systematic variation in latency across the vowel continua. Thus, the conclusion of Chistovich et al. that imitation bypasses conscious response selection, regardless of delay, is not contradicted.

Some striking differences among conditions are evident in Fig. 1. Both subjects showed consistently longer reaction times for immediate than for delayed (+750 ms) imitation. (Note that the delayed-imitation latencies were measured from the onset of the "go" signal; to obtain stimulus-response latencies, 750 ms must be added to the times shown.) This is not unexpected: After all, the subjects knew exactly what to produce when the "go" signal occurred in the delayed condition, whereas in the immediate condition they had no such advance knowledge, and therefore had to wait until at least part of the stimulus vowel had been processed. It should also be noted that the latencies of subjects DW and BR were much longer than those of L.C. Neither of the present subjects was capable of the very close shadowing evinced by L.C. [2]. The fact that all responses were initiated from a closed mouth position may have something to do with this; L.C.'s articulatory resting position was not reported.

In the deferred imitation condition, there was increased temporal uncertainty about the moment of occurrence of the "go" signal; hence, an increase in reaction times might have been expected relative to the delayed imitation condition. Such an increase was shown by BR but not (or to a much lesser extent) by DW. The reason for this difference between subjects is not clear. Another difference of uncertain origin is DW's slower response to the [i]-[æ] continuum.

3.2. *Formant frequencies*

Presentation of formant frequency data is simplified by the finding that two factors seemed to play only a very minor role. First, formant frequencies remained roughly constant or fell slightly over the course of the response vowels. Frequency measures at response vowel onset and two thirds into the vowel followed patterns almost identical to those of the mean formant frequencies across the whole vowel; therefore, the latter were chosen as the primary dependent variable. Second, formant frequencies were virtually identical across the three imitation conditions. Therefore, mean formant frequencies will be presented averaged across conditions ($n = 36$ per stimulus).

Figure 2 plots the formant frequencies of the two subjects' responses to the 12 members of each vowel continuum. The formant frequencies of the stimulus vowels are indicated by the dashed lines. These plots are modelled after Figure I.-A-4 of Chistovich et al. [2], which showed several steps, especially in the function for *F*2. No such steps (i.e., true plateaus) can readily be discerned in Fig. 2, except at the [u] end of the [u]-[i] continuum for DW. The response formant frequencies are not a linear function of the stimulus formant frequencies, however. Changes in the slope of the *F*2 function are evident especially along the [u]-[i] continuum, and also at the [i] end of the [i]-[æ] continuum for DW. Note also that the [i] stimulus was not well matched to the subjects' vocal capabilities, even though it was based on Peterson and Barney's [9] male norms. For both DW and BR, *F*2 and especially *F*3 frequencies in the response vowels were much lower than in [i]-like stimuli.

A closer examination of these stimulus-response relationships is possible in Figs. 3 and 4, which present two-dimensional formant frequency plots. Figure 3 shows the data for the [u]-[i] continuum in *F*2-*F*3 space; *F*1 varied very little across this series. Each stimulus vowel is connected by a line to its corresponding response. There is an obvious similarity between the two subjects' data in that the linear stimulus continuum is mapped onto a compressed, curvilinear contour in *F*2-*F*3 space. It is also evident that there are local expansion and compression effects that are quite different for the two subjects. DW did not distinguish among stimuli 1–3; this is the only clear instance of a categorical response in the present data. His responses to stimuli 3–6

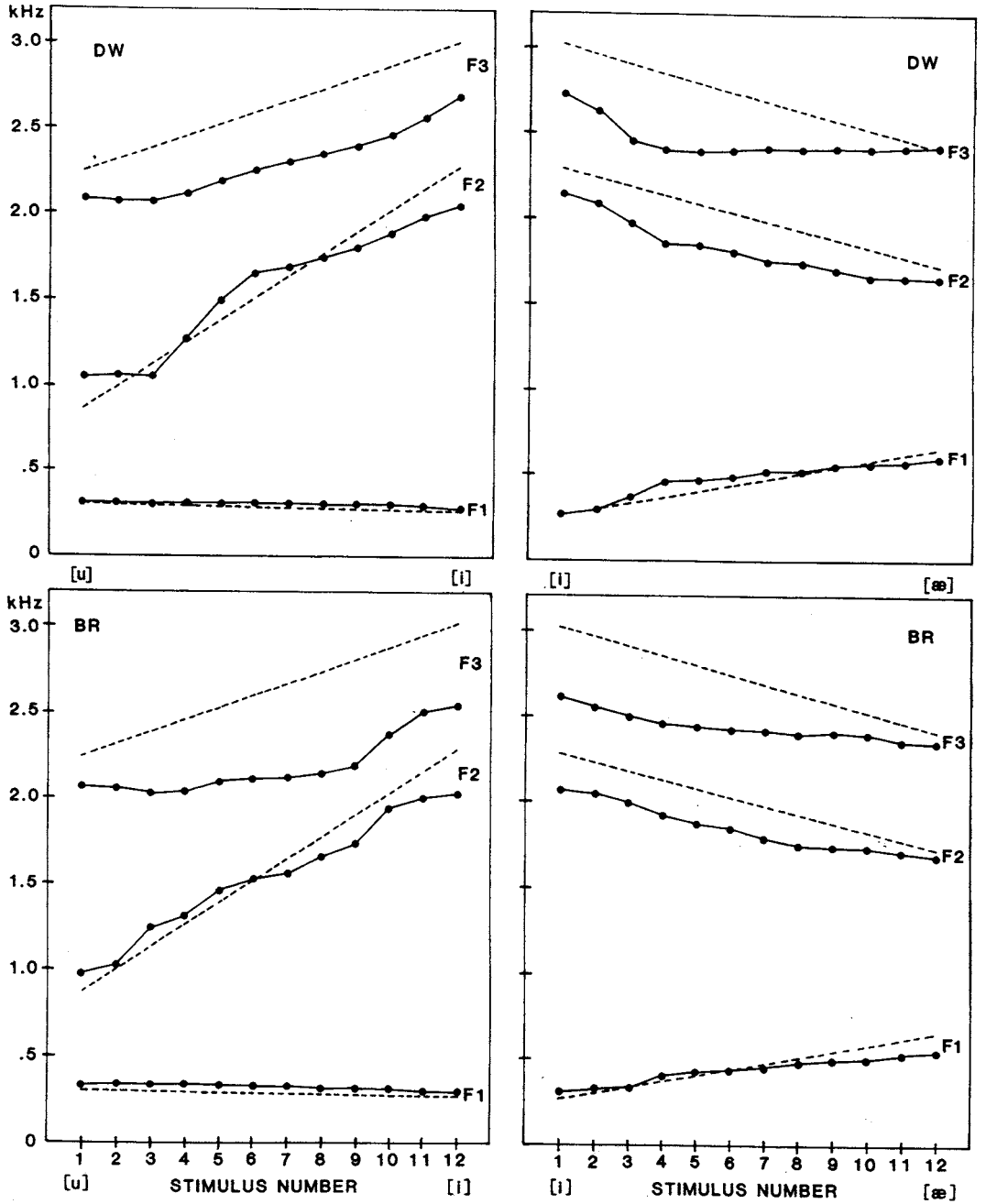

Fig. 2. Response formant frequencies (averaged over 36 responses) as a function of stimulus number, vowel continuum, and subject. The dashed lines represent the stimulus formant frequencies.

were spaced widely apart, while stimuli 6–9 were mapped onto a compressed response space; responses to stimuli 9–12 roughly matched the distances between stimuli. Stimuli 6–9 could be interpreted as forming a second category on this continuum, although DW clearly was capable of responding discriminatively within that category. The stimulus-response mapping for BR is quite different. Major gaps appear between responses to stimuli 2–3, 4–5, and especially 9–10. Thus, four clusters of responses may be distinguished, perhaps corresponding to categories of some kind. Clearly, however, BR was able to respond distinctively to stimuli within each of these categories.

It might be added that the stimulus-response functions for *F*1, which are difficult to discern in Fig. 2, were quite systematic and different for the two subjects, even though *F*1 changed over only a 30 Hz range. For DW, *F*1 was practically constant for responses to stimuli 1–10 and then decreased rapidly. For BR, on the other hand, *F*1 was roughly constant for stimuli 1–4 and then decreased almost continuously, although a local increase for stimulus 9 might lead one to consider stimuli 5–9 as a second grouping. Such a grouping would be consistent with the patterning seen in the lower panel of Fig. 3.

The data for the [i]-[æ] continuum are shown in Fig. 4; they are plotted in *F*1-*F*2 space, disregarding *F*3. Again, the response trajectories in this space are similar for the two subjects; they are very nearly linear, parallel to the stimulus continuum, and compressed, although more so for BR. The linearity reflects the fact that, as in the stimuli, changes in *F*1 and *F*2 were correlated in the responses. The correlation (computed between the *F*1 and *F*2 frequency differences of responses to adjacent stimuli on the continuum) was 0.97 for DW and 0.81 for BR ($p < .001$). The exact stimulus-response mapping, however, again shows individual differences. DW has some striking gaps between responses to stimuli 2–3–4; responses to stimuli 4–5, 7–8, and 10–11, on the other hand, are very similar. Clearly, there are strong distortions in this mapping, but they do not reveal a clear categorical structure. The same can be said for BR's data, although the distortions are less strong here, with the largest gap occurring between responses to stimuli 3–4. Neither subject's *F*3 values lead to different conclusions, as is evident from Fig. 2: For DW, *F*3 changed rapidly in response to stimuli 1–4 and then remained completely insensitive to stimulus variations; for BR, similarly, *F*3 decreased more rapidly at first and then decreased more gradually from stimulus 4 on. There are no indications of any local categories in these *F*3 tracks.

To summarize, these formant frequency data offer ample evidence for nonlinearities in stimulus-response mapping, but little evidence for response categories in the strict sense. Moreover, they offer little support for Kent's [4] tentative conclusion that responses to the [i]-[æ] continuum are more categorical than those to the [u]-[i] continuum. We turn now to an examination of the formant frequency standard deviations.

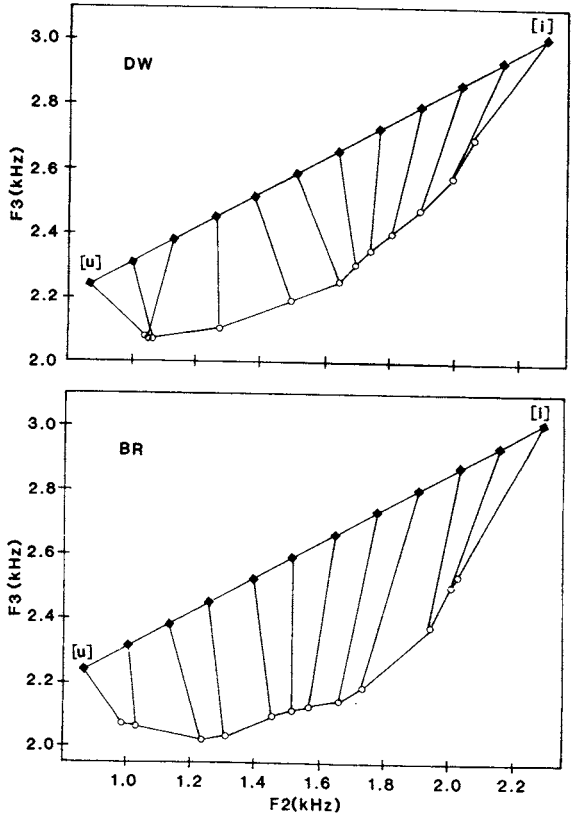

Fig. 3. Average responses (circles) to stimuli from the [u]-[i] continuum (diamonds) in *F*2-*F*3 space. Corresponding stimuli and responses are connected by straight lines.

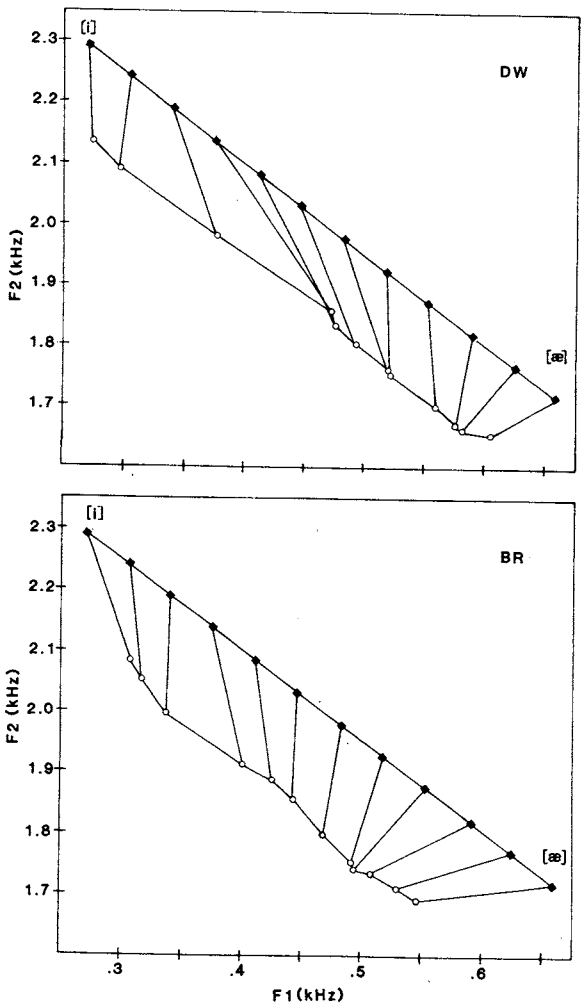

Fig. 4. Average responses (circles) to stimuli from the [i]-[æ] continuum (diamonds) in *F1*-*F2* space.

3.3. Standard deviations

To simplify presentation, the standard deviations, like the formant frequencies, will be presented averaged across the three imitation conditions. These average within-condition standard deviations do not include between-condition variability, but this variability was small, as noted above. The absolute magnitude of response variability was comparable across conditions. The patterns of standard deviations, on the other hand, showed considerable differences among conditions. These differences must be viewed with caution, however, because of the similarity in mean formant frequencies across conditions; moreover, we have no estimate of measurement error (i.e., of the random variability of standard deviations). Some differences will be mentioned below.

The standard deviations for all three formants are displayed in Fig. 5, separately for the two subjects and the two stimulus continua. Some general observations may be made at the outset: For both subjects, the standard deviations of *F1* are larger on the [i]-[æ] continuum than on the [u]-[i] continuum, perhaps because only the former requires active control of degree of jaw opening. The standard deviations of *F2*, on the other hand (and, to some extent, those of *F3*), are much larger and more variable on the [u]-[i] continuum. This probably reflects the relative unfamiliarity of the vowel sounds along that continuum (cf. [4]). The fact that *F3* variability is often lower than *F2* variability may also be noted.

If there is a quasi-categorical structure underlying the imitation responses, then the standard deviations should increase whenever the slope of the formant frequency function increases (i.e., in the "between-category" regions). The striking peak in the *F2* function for DW on the [u]-[i] continuum indeed coincides with the rapid change in response *F2* frequency between stimuli 3 and 6 (see Fig. 2). The peaks at stimuli 4 and 9 in the *F2* function for BR, on the other hand, have no such clear correlate in the pattern of mean *F2* frequencies (Fig. 2). A finding not shown in Fig. 5 is that, for both subjects, the *F2* standard deviation peaks along the [u]-[i] continuum were most pronounced in the immediate imitation condition. The patterns of *F1* and *F3* standard deviations show only occasional correspondence to the pattern of *F2* standard deviations, and they also varied considerably across conditions. (The large standard deviation of *F3* for the [i] endpoint stimulus on both stimulus continua for subject DW was entirely due to the delayed imitation condition, for unknown reasons; hence the parentheses in Fig. 5.)

Along the [i]-[æ] continuum, a correlation of *F1* and *F2* standard deviations can be seen for DW ($r = 0.70$, $p < .01$) but not for BR ($r = -0.03$). Standard deviation peaks for *F1* seemed most pronounced in the immediate imitation condition. The pattern of *F1* and *F2* standard devia-

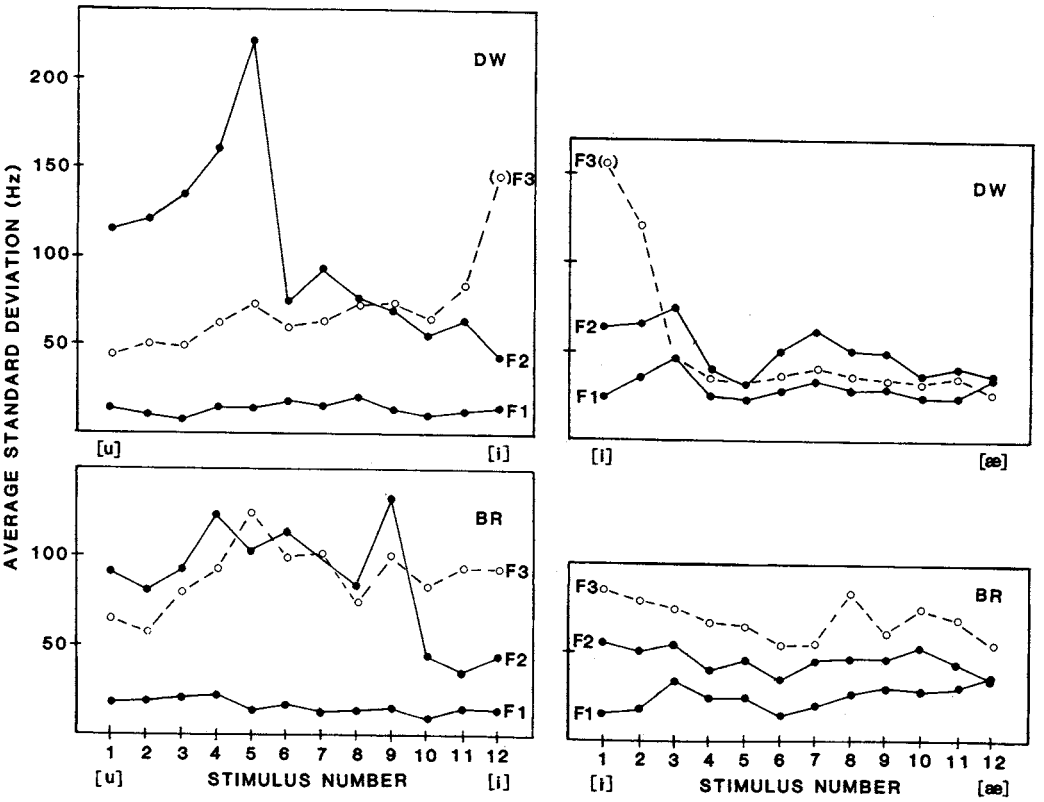

Fig. 5. Formant frequency standard deviations (averaged over imitation conditions) as a function of stimulus number, vowel continuum, and subject.

tions finds some correspondences in slope changes in the formant frequency functions (Fig. 2). Correlations between the standard deviations and the absolute frequency differences between responses to stimuli n and $n + 1$ on the [i]-[æ] continuum were significant for subject DW, both for $F1$ ($r = 0.80$, $p < .01$) and for $F2$ ($r = 0.74$, $p < .01$). For BR, on the other hand, the standard deviations were not significantly correlated with the pattern of response formant frequencies.

3.4. Frequency distributions

As might have been expected from the similarity in mean formant frequencies across imitation conditions, the formant frequency distributions for each continuum were highly similar across conditions also. Therefore, histogram envelopes will be presented for responses in all three conditions combined ($n = 432$ per continuum). They are shown in Figs. 6 and 7.

The data for the [u]-[i] continuum appear in Fig. 6. The $F1$ distributions at the bottom are strongly unimodal and reflect the general upward shift of $F1$ in the subject's responses, relative to the stimuli. The $F2$ distributions in the middle panels, on the other hand, cover the stimulus range rather well and show evidence of trimodality for both subjects. The first and third peaks of both subjects are located similarly near the endpoints of the continuum and presumably represent /u/ and /i/ categories, respectively. The middle peak is located differently: closer to [i] for DW but in the center of the continuum for BR. The clear trimodality of these distributions is surprising after the somewhat ambiguous patterns of the formant frequency and standard deviation curves. It indicates definite response preferences (or avoidances?) on the part of the subjects, although it should be noted that the speakers were

able to produce an *F2* frequency along the continuum, except for that corresponding to the [i] endpoint stimulus. The *F2* response distributions for individual stimuli (not shown here) were also examined for bimodality. Although there were several distributions with two peaks (e.g., stimuli 4 and 5 for DW, stimuli 3 and 9 for BR), these peaks generally did not coincide with the major peaks seen in the overall histogram. They might indicate tendencies to avoid certain *F2* values, or else they reflect just random variability. The *F3* distributions (top panels) are strongly skewed toward low frequencies, and considerable "undershoot" is present, as noted earlier. In addition to the major peak reflecting the constancy of *F3* over part of the continuum (cf. Fig. 2), two minor peaks may be distinguished for each subject; however, their locations are only partially consistent with those of the *F2* peaks.

The histogram envelopes for the [i]-[æ] continuum are plotted in Fig. 7. The *F1* distributions at the bottom show good coverage of the stimulus range as well as very pronounced peaks, four for DW and three for BR. The peaks correspond to different stimuli for the two subjects. The *absolute* frequency locations of the three major peaks, however, are remarkably similar: 325, 475, and 550 Hz (± 12 Hz) for DW; 324, 468, and 530 Hz (± 7 Hz) for BR. There is much less correspondence in the *F2* distributions for the two subjects (middle panels). For DW, the *F2* histogram is clearly trimodal: One mode is in the [i]-[I] region

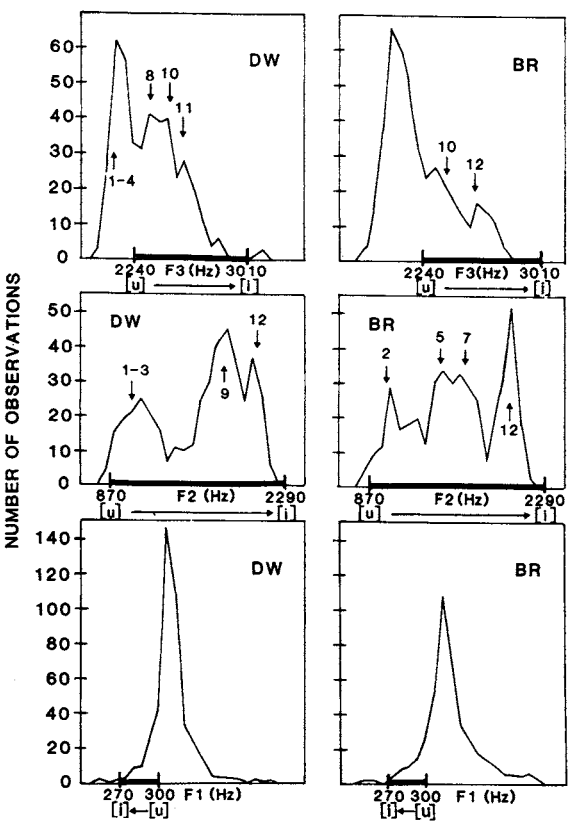

Fig. 6. Formant frequency distributions for responses to stimuli along the [u]-[i] continuum. The heavy line at the bottom of each graph indicates the range of formant values along the stimulus continuum. Numbered arrows indicate stimuli for which the mean formant frequencies of the responses fell in the vicinity of peaks in the histogram. Note that the distributions for the different formants are not aligned with respect to each other or across subjects; rather, they are spread out over a constant number of histogram bins.

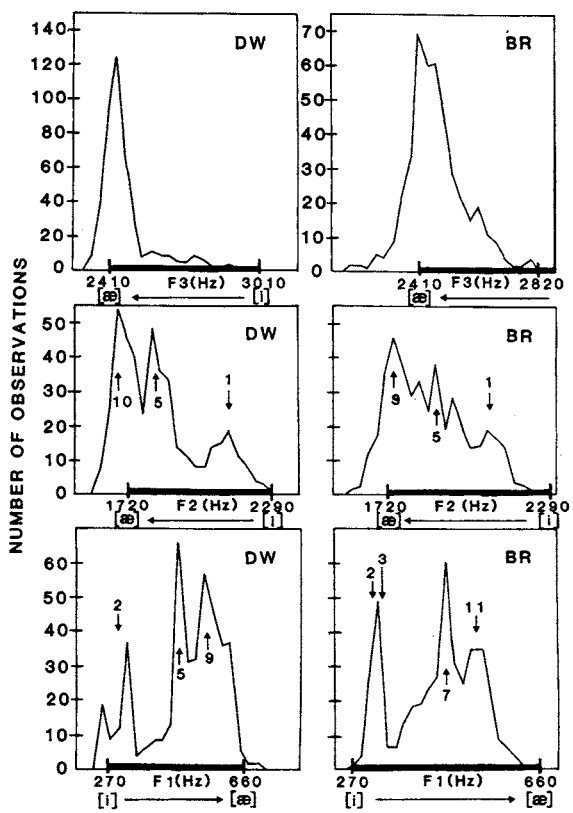

Fig. 7. Formant frequency distributions for responses to stimuli along the [i]-[æ] continuum.

and two modes are in the [ɛ]-[æ] region. The distribution for BR is less clear, having five peaks. Note the general asymmetry of these distributions; there seemed to be a strong "pull" toward lower *F*2 frequencies in the subjects' responses, apart from the absolute downward shift in *F*2. The *F*2 distributions for some individual stimuli (not shown) showed signs of bimodality (e.g., stimuli 3 and 8 for DW, stimuli 3 and 9 for BR), but not so strongly as to suggest discrete response categories. As on the [u]-[i] continuum, the whole range of *F*2 frequencies was represented in the responses. The *F*3 distributions (top panels) were unimodal and were centered at the lower end of the stimulus range for both subjects (cf. Fig. 2).

These data may be summarized by stating that (1) subjects show pronounced preferences for particular formant frequencies in their responses to both stimulus continua, and (2) the two subjects' responses are rather similar with regard to *F*1 and *F*3, with individual differences residing primarily in the *F*2 distributions, particularly in responses to stimuli from the centers of the continua.

3.5. *Written identification*

The subjects' phonemic labeling responses to the stimuli on the [i]-[æ] continuum are shown in Fig. 8. It is evident that DW and BR applied somewhat different criteria, possibly because of their different language backgrounds. DW divided the continuum fairly consistently into five categories: /i/ (stimuli 1–4), /I/ (5), /e/ (6), /ɛ/ (8–11), and /æ/ (12). Stimulus 7 was ambiguous (i.e., less than 75 percent responses in any category). BR, on the other hand, applied only four categories consistently: /i/ (1–2), /e/ (4–5), /ɛ/ (7–9), and /æ/ (11–12). Stimuli 3, 6 and 10 were ambiguous to this listener. The /I/ category was not consistently applied by him to any stimulus; stimuli 2–7 received a few responses in that category. Essentially, for BR the /e/ category occupied the place of DW's /I/ and /e/ categories. A prominent role of /e/ might be expected in a native speaker of German, which has a monophthongal /e/ phoneme; however, the /I/ phoneme is also distinctive in German, not to speak of BR's long exposure to English. It will also be noted that BR's phoneme boundaries are all shifted toward the lower end of the continuum relative to DW's boundaries. The reason for this shift is not immediately evident.

The patterns of these labeling responses may be compared with the histogram peaks in Fig. 7. For DW, the three major peaks in the *F*1 distribution (bottom left) may be identified with the /i/, /I/, and /ɛ/ categories, respectively. There are no peaks corresponding to /e/ and /æ/. The three peaks in the *F*2 distribution for DW (center left) correspond to /i/, /I/, and /ɛ/ again. The three peaks in BR's *F*1 distribution (bottom right) correspond somewhat less clearly to /i/, /ɛ/, and /æ/, while the three major peaks in his *F*2 distribution (center right) reflect more unambiguously /i/, /e/, and /ɛ/. There are no peaks for /e/ in *F*1 and for /æ/ in *F*2. For DW, and to some extent also for BR, it appears that the number of functional categories on this continuum is smaller in production than in perception, contrary to the conclusions of Chistovich et al.[2].

This should not be taken to imply, however, that the subjects were somehow less sensitive to stimulus differences in the imitation task than in

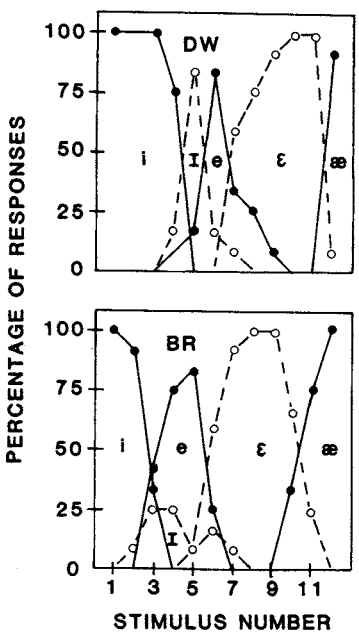

Fig. 8. Phonemic identification responses to stimuli along the [i]-[æ] continuum.

the phonemic labeling task. On the contrary, Fig. 4 shows clearly that DW, for example, imitated quite differently the three stimuli (1–3) uniformly labeled as /i/. In fact, the clustering of modal responses in *F*1-*F*2 space (Fig. 4) is difficult to relate to the phonemic category boundaries (Fig. 8), which leaves the issue of the origin of the categorical tendencies in production unresolved.

Phonemic labeling of vowels does not reflect the extent of the subjects' perceptual sensitivity, as is abundantly clear from many earlier categorical perception studies (see [11] for a review). This is also demonstrated by the absolute identification responses of the present two subjects, which are graphed in Fig. 9. These responses are a monotonic function of stimulus number and show good discrimination of most adjacent stimuli, especially toward the ends of the continua. Thus, for example, it is clear that DW not only imitated differentially stimuli 1–3 on the [i]-[æ] continuum (all labeled /i/) but also was able to discriminate them perceptually without special training. More interesting, perhaps, is the observation that he also discriminated stimuli 1–3 on the [u]-[i] continuum, which he had imitated in identical fashion (cf. Fig. 3). Whether this implies a limitation on production or a perceptual limitation caused by the higher stimulus uncertainty in the imitation task remains to be seen. The poorer discrimination in the centers of the continua may be a consequence of the AXB paradigm, which provided endpoint anchors that facilitated discrimination of stimuli in their vicinity. The steps in the functions bear only a vague correspondence to the compression regions in the response formant space (Figs. 3 and 4).

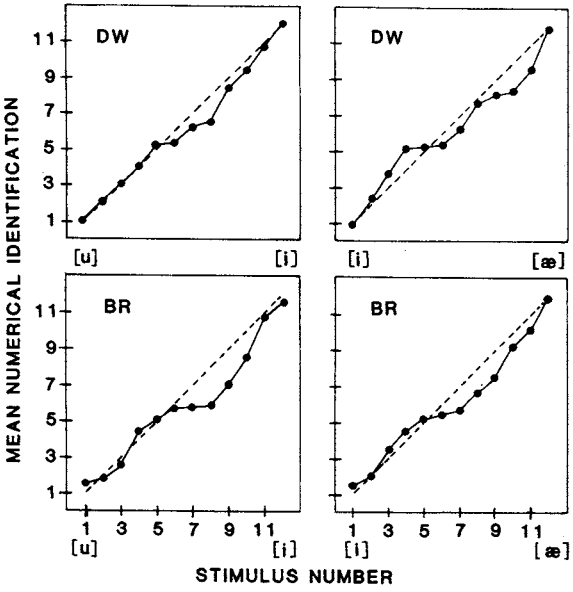

Fig. 9. Average numerical identification responses to stimuli along both vowel continua. The dashed lines have a slope of 1.

4. Summary and conclusions

The data presented here are preliminary and do not permit any strong conclusions, especially since their pattern exhibits some of the ambiguities observed by Kent [4], after whose study the present experiment was modeled. A few tentative observations can be made, however, which should help guide future research on this topic.

(1). It appears from the present data that the pattern of vowel imitation responses is essentially insensitive to the delay between stimulus and response (from 300 to 3000 ms). This agrees with the conclusions of Chistovich et al. [2], although it should be noted that the present subjects did not shadow as rapidly as subject L.C. Whatever internal representation of the stimulus mediates vocal imitation, it seems to be both (virtually) immediately available and relatively long-lasting in memory. A trend toward more pronounced patterns of formant variability in immediate imitation was observed.

(2) Vocal response latencies do not seem to vary much across a vowel continuum, as also observed by Chistovich et al. [2]. Thus, imitation seems to bypass conscious response selection, regardless of delay.

(3) In contrast to Chistovich et al. [2] and, to some extent, in contrast to Kent [4], we found no discrete steps in the response formant frequency functions. There was ample evidence, however, for nonlinearities in the stimulus-response relationships. The pattern of response variability resembled that of the formant frequencies for one subject only; basically, however, changes in mean formant frequencies and standard deviations can be assumed to reflect the same underlying tenden-

cies. The absence of a strictly categorical structure in responses to isolated, stationary vowels is in agreement with the categorical perception literature (see [11]). It is not clear whether the stimulus-response nonlinearities should be interpreted in terms of perceptual categories, especially since they were difficult to relate to the phonemic labeling responses. The hypothesis needs to be pursued that these distortions have an independent origin in the production system.

(4) Response frequency histograms for individual formants showed very pronounced peaks and valleys. Thus the subjects had definite response preferences, which seemed to correspond to some of their phonemic categories. Not all phoneme categories were represented, however, and some categories had a peak only in a single formant. The relation of the histogram peaks to the stimulus-response nonlinearities is not easily characterized, although they are, of course, not independent. Generally, peaks correspond to regions of compression in the response formant space, and valleys correspond to regions of expansion. The response histograms, however, seem to provide a much clearer indication of underlying categories than do the formant frequency plots.

(5) We have found little support for Kent's [4] tentative claim that responses to the [u]-[i] continuum, which does not harbor familiar phonemic categories apart from the endpoints, are less categorical than responses to the [i]-[æ] continuum, which spans five phonemic categories in English. The *F*2 histograms show three major peaks on both continua. It is true, however, that on the [i]-[æ] continuum *F*1 histograms provide additional striking information about response preferences, whereas *F*1 along the [u]-[i] continuum is too restricted in range to be informative. Response variability, especially in *F*2, was also much larger on the [u]-[i] continuum, presumably due to the unfamiliarity of the vowel sounds on this continuum.

(6) There were considerable individual differences between the two subjects, which might be related to their different language experiences. Some instances of congruity were also noted. Individual differences seemed most pronounced in the pattern of *F*2 frequencies.

(7) Chistovich et al. [2] hypothesized the existence of an intermediate stage of representation, characterized by a number of categories exceeding (but including) the functional categories in the subjects' language. With regard to the [u]-[i] continuum, the hypothesis is supported, since both subjects seemed to have an additional category between the two familiar endpoints. The hypothesis is not supported for the [i]-[æ] continuum, however, where the number of categories in production (i.e., peaks in the formant histograms) was smaller than that of relevant vowel phonemes in the language. It is possible that this continuum was not sampled finely enough to reveal its full categorical structure in vocal reproduction.

In conclusion, the present results leave unresolved the issue of whether an intermediate mental representation needs to be postulated to account for nonlinearities in vowel imitation responses. These nonlinearities exist, however, and the search for their origin should continue. We plan to extend our research in several directions: by testing monolingual speakers of different languages to examine the role of linguistic experience; by matching the stimuli more closely to the subjects' vocal capabilities (as done originally by Chistovich et al., [2]) so as to eliminate distortions that may arise in perceptual normalization; and by studying in more detail the possibility that the observed nonlinearities have their origin in articulation itself. Eventually, we also want to use more realistic, time-varying speech stimuli. In general, the thrust of our research will be to disentangle the perceptual and articulatory factors that jointly constrain vocal imitation. This enterprise seems interesting and worthwhile, and it provides but one example of the immense stimulus to speech research provided by the pioneering work of Ludmilla Chistovich and her colleagues.

Acknowledgments

This research was supported by NICHD Grant HD-01994 and BRS Grant RR-05596 to Haskins Laboratories. We thank Michael Studdert-Kennedy and Ignatius Mattingly for their helpful comments.

References

[1] L.A. Chistovich, "Relation between speech production and speech perception", in: M.P.R. van den Broecke and A. Cohen, eds., *Proceedings of the Tenth International Congress of Phonetic Sciences*, Foris Publications, Dordrecht, 1984, pp. 55–58.

[2] 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.

[3] L.A. Chistovich, G. Fant and A. de Serpa-Leitão, "Mimicking and perception of synthetic vowels, part II", *Quarterly Progress and Status Report* (Royal Technical University, Speech Transmission Laboratory, Stockholm), No. 3, 1966, pp. 1–8.

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

[5] R.D. Kent, "Auditory-motor formant tracking: A study of speech imitation", *J. Speech and Hearing Res.*, Vol. 17, 1974, pp. 203–222.

[6] R.D. Kent, "Imitation of synthesized vowels by preschool children", *J. Acoust. Soc. Am.*, Vol. 63, 1978, pp. 1193–1198.

[7] R.D. Kent, "The imitation of synthesized English and non-English vowels by children and adults", *J. Psychol. Res.*, Vol. 8, 1979, pp. 43–60.

[8] R.D. Kent and L.L. Forner, "Developmental study of vowel formant frequencies in an imitation task", *J. Acoust. Soc. Am.*, Vol. 65, 1979, pp. 208–217.

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

[10] D.B. Pisoni, "Auditory and phonetic memory codes in the discrimination of consonants and vowels", *Perception and Psychophysics*, Vol. 13, 1973, pp. 253–260.

[11] B.H. Repp, "Categorical perception: Issues, methods, findings", in: N.J. Lass, ed., *Speech and Language: Advances in Basic Research and Practice*, Vol. 10, Academic Press, New York, 1984, pp. 243–335.

[12] M.E.H. Schouten, "Imitation of synthetic vowels by bilinguals", *J. Phonetics*, Vol. 5, 1977, pp. 273–283.