Perceptuomotor bias in the imitation of steady-state vowels\textsuperscript{a)}

Gautam K. Vallabha\textsuperscript{b)} and Betty Tuller

Center for Complex Systems and Brain Sciences, Florida Atlantic University, 777 Glades Road, Boca Raton, Florida 33431

(Received 11 July 2003; revised 20 April 2004; accepted 3 May 2004)

Previous studies suggest that speakers are systematically inaccurate, or biased, when imitating self-produced vowels. The direction of these biases in formant space and their variation may offer clues about the organization of the vowel perceptual space. To examine these patterns, three male speakers were asked to imitate 45 self-produced vowels that were systematically distributed in F1/F2 space. All three speakers showed imitation bias, and the bias magnitudes were significantly larger than those predicted by a model of articulatory noise. Each speaker showed a different pattern of bias directions, but the pattern was unrelated to the locations of prototypical vowels produced by that speaker. However, there were substantial quantitative regularities: (1) The distribution of imitation variability and bias magnitudes were similar for all speakers, (2) the imitation variability was independent of the bias magnitudes, and (3) the imitation variability (a production measure) was commensurate with the formant discrimination limen (a perceptual measure). These results indicate that there is additive Gaussian noise in the imitation process that independently affects each formant and that there are speaker-dependent and potentially nonlinguistic biases in vowel perception and production. © 2004 Acoustical Society of America. [DOI: 10.1121/1.1764832]

PACS numbers: 43.71.Es, 43.71.An [RD]

Pages: 1184–1197

I. INTRODUCTION

A persistent debate in speech research is over the nature of vowel representation and memory and over which methodologies constitute appropriate measures of the vowel percept. Vowel perception has several phenomena in common with nonspeech perception, most prominently context-sensitivity and noncategoricality (Pisoni, 1973). Consequently, auditory processes are thought to influence vowels more than consonants. On the other hand, vowel discrimination is partly mediated by phonetic labels (Repp et al., 1979) and is worse among more prototypical exemplars of a category (Iverson and Kuhl, 1995; Shigeno, 1991). There is also a significant stimulus order effect in vowel discrimination that seems to depend on the phonetic range and location of the tokens (Repp and Crowder, 1990). Thus the perception of a vowel is influenced by its auditory quality, its phonetic status, and its location in the vowel space, and the tension among these three aspects has not been experimentally resolved.

One source of the experimental tension is that the standard methods of evaluating the vowel percept—categorization, identification, AX or ABX discrimination, multidimensional scaling and goodness ratings—share many of the same limitations. First, they quantize what is essentially a continuous and graded percept. Second, discrimination and multidimensional scaling methods are very time-consuming. The tradeoff between stimulus resolution and the number of stimuli restricts a discrimination experiment to only a small part of the vowel space (a surprisingly large number of results are based upon the one-dimensional [i–e–æ] continuum). Consequently, large areas of the vowel space are as yet unexplored. Finally, a discrimination experiment can only examine changes in phonetic quality that are projected onto the stimulus continuum; for instance, the only context effects observable with one-dimensional continua are contrast and assimilation. One way to avoid these problems is to exploit the fact that listeners are also speakers and use vowel imitation as a measure of vowel perception, and the F1/F2 patterns of the imitations as clues to vowel organization. Vowel production is intrinsically continuous and multidimensional, so in principle a subject’s imitation can convey subcategorical qualities of the vowel percept.

There is substantial evidence that imitation is deeply linked to speech perception and production. Infants only a few weeks old attempt to mimic the vocalizations of adults around them and 20-week old infants can imitate the point vowels [i, a], and [u] (Kuhl and Meltzoff, 1996). Such imitations, along with spontaneous babbling, are thought to forge the perception–production link that allows more advanced articulations (Kuhl, 2000). Moreover, when adult subjects are asked to listen to a syllable sequence and repeat it with as little delay as possible, they perform with remarkable accuracy at latencies as short as 150 ms (Porter and Lubker, 1980; also see Fowler et al., 2003). In fact, the response latency for shadowing is smaller than the latency for a simple response (i.e., detecting the vowel and uttering a standard response). These results suggest a fast subcognitive link between speech perception and production. This link may be covertly active even in regular speech: Pardo and Fowler (2000) report that the productions of two speakers are more similar to each other after they had conversed than

\textsuperscript{a)}This work was part of the first author’s Ph.D. dissertation at the Center for Complex Systems and Brain Sciences, Florida Atlantic University. Portions of this work were also presented at the 142nd meeting of the Acoustical Society of America [J. Acoust. Soc. Am. 110, 2657(A) (2001)].

\textsuperscript{b)}Author to whom correspondence should be addressed. Currently affiliated with the Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA. Electronic mail: vallabha@cnbc.cmu.edu
before (without any instructions to imitate). Imitation has also been posited to be the driving force behind sound change in language (de Boer, 2000). For these reasons, imitation seems a particularly apt method for studying speech in general and the vowel space in particular.

The utility of imitation to speech perception research has been recognized since the 1960s. In one of the first imitation studies, Chistovich et al. (1966) synthesized 12 vowels along an [a–e–i] trajectory in F1/F2 space and asked a phonetically trained female subject to imitate them. When the mean F2 of the imitations was plotted against the ordinal of the stimulus (1–12), the resulting curve showed four well-defined plateaus. In addition, the F2 histogram showed four peaks near the category centers and the standard deviation of F2 showed peaks near the boundaries of the putative categories. Chistovich et al. interpreted these results to indicate that the subject had between four and six categories across the [a–e–i] continuum. Since the subject’s native language (Russian) has only 3 vowel phonemes across the continuum, they posited that vowel representation is fundamentally discrete but at a finer granularity than the native phonemic categories. Kent (1973) failed to replicate the sub-phonemic granularity with American English speakers. Repp and Williams (1985) had a little more success: They asked two phonetically trained male speakers to imitate 150 ms synthetic vowels along the [æ–i] and [u–i] continua. There were clear peaks in the resulting formant histograms, indicating that the speakers had distinct response preferences. However, the formant frequency curves did not exhibit the plateaus observed by Chistovich et al. and the standard deviations did not show any consistent pattern across the two speakers.

It should be noted that both Kent (1973) and Repp and Williams (1985) found large intersubject differences in formant patterns. In retrospect, these differences are to be expected, since the vowels produced by a speaker are influenced by the physiology of the speaker and idiosyncrasies in the manner of production (Johnson et al., 1993). These differences can in turn interact with how vowels are perceived by that speaker (Fox, 1982). Thus, each speaker has a different vowel space and, potentially, a different pattern of imitation responses. The intersubject differences also imply that (a) it is methodologically risky to average formant data across subjects and (b) response preferences in a speaker’s imitations of synthetic vowels may simply be due to a mismatch between the synthetic targets and the production capabilities of the speakers.

Repp and Williams (1987) tested the “stimulus mismatch” explanation by asking speakers to imitate self-produced rather than synthetic vowels. They used as subjects the same two speakers from their earlier experiment. For each speaker, they chose 12 self-produced targets that were approximately equidistant along the [u–i] and [æ–i] continua. In addition, they also recorded the speakers’ [hVd] utterances, and were thus able to map the speaker’s “prototypical” vowels. Three of the results are striking: (1) The variability of the self-imitations was remarkably similar to the imitations of synthetic vowels; (2) the self-imitations were consistently inaccurate, regardless of the latency of the imitation; (3) the imitations seemed to be pulled towards regions with more prototypes, but seemed to end up between prototypes rather than at the prototypes. These results indicate that the imitation inaccuracies are not due simply to a mismatch between the stimulus and the speaker’s production capabilities. Moreover, the similarity between the imitations of synthetic and self-produced vowels suggests that the inaccuracies are shaped by the phonetics of the targets and not just their auditory qualities.

In summary, the line of research initiated by Chistovich et al. (1966) has shown that speakers prefer certain formant frequency regions. There are also clear nonlinearities in vowel imitation that hint at the existence of representational categories but none of the studies have been able to identify the putative categories or relate them to phonemes or allophones.

One other inspiration for the study reported here comes from the imitation of visual stimuli. Studler et al. (1991) created a grid of dots on a rectangular sheet of paper and presented them one at a time to each subject, who then had to reproduce the location of the dot from memory. The reproductions of the dots’ locations were systematically biased toward the corners of the sheet, which Studler et al. interpreted as evidence for a Gestalt structure in the visual field, i.e., that the visual system organized itself with respect to the boundary of the paper and distorted the perception of the homogenous stimulus array. Wildgen (1991) proposed that vowel systems may be organized in an analogous manner, with the corner vowels [i], [a], and [u] functioning as the attractors. This proposal is plausible since there is abundant evidence that the perceptual space of vowels is warped even with respect to an “auditory” space of bark or mel units, and that the warping is influenced by native vowel categories (e.g., Kewley-Port and Atal, 1989; Iverson and Kuhl, 1995).

In the present experiment, we reexamine the issue of vowel organization using a self-imitation paradigm. Our motivations are twofold. The first is to test whether the complex imitation patterns seen with one-dimensional continua such as [u–i] and [æ–i] would become more coherent with two-dimensional (2D) stimulus grid (and if they do, whether they are influenced by the natural vowels of the speaker). The second motivation is to test the null hypothesis that inaccuracies in imitation are caused by random articular or perceptual fluctuations and do not reflect any deeper phonetic organization. To evaluate this claim, we compare the speakers’ patterns of imitation inaccuracies with those from an articulatory model (Rubin et al., 1981) and a perceptual model based on formant difference limens.

II. METHOD

The experiment was conducted over two days. On Day 1, subjects were asked to imitate a set of 100 synthetic vowel-like stimuli; the purpose of this step was to encourage the subject to produce vowel-like sounds that were likely to be well distributed in formant space. They were also asked to read [hVd] words in citation and sentence contexts. For each subject, 45 vowels were chosen from the 100 self-produced vowels to serve as the targets. On Day 2, the 45 self-produced targets were presented for imitation. The steps are described in more detail below.
A. Subjects

The subjects (CD, DR, FC) were three male native speakers of American English who volunteered to participate in the experiment. CD (39 years old) grew up near New York City; DR (40) grew up in North Carolina until the age of 14 years and in New York City after that; FC (31) grew up in New Hampshire. The subjects did not have any phonetic training, were not fluent in a second language, and did not have any hearing problems. FC and CD had participated in previous imitation experiments.

B. Stimuli

One hundred steady-state vowel-like stimuli were synthesized using the 1988 version of the Klatt synthesizer (KLSYN88a, Sensimetrics Corporation, Cambridge, MA). Each stimulus had three formants, with F1 taking on one of 8 values (300–750 Hz in steps of ~65 Hz), F2 taking on one of 13 values (998–2400 Hz in steps of ~115 Hz), and F3 constant at 2500 Hz [Fig. 1(a)]. For all stimuli, F0 was 120 Hz and the duration was 200 ms. F0 and the amplitude of voicing were linearly ramped at the beginning and end of each stimulus for 30 ms. The bandwidths of the formants were 60 Hz (F1), 90 Hz (F2) and 150 Hz (F3). Finally, the stimuli were sampled at 10 kHz.

C. Procedure

The subject was seated in a sound-insulated booth, and the vowel stimuli were presented binaurally over headphones (Telephonics TDH-39P) at a comfortable loudness. The subject’s productions were recorded using a microphone (Sony electret condenser, ECM-23F) and digitized at a sampling rate of 10 kHz. The microphone was mounted on a table in front of the subject, and its position was adjusted so that it was about 10 inches from the subject’s mouth when the subject was comfortably seated. Both the stimulus presentation and recording were computer-controlled using the Desklab system (Model 216; Gradient Technology, Inc., now defunct).

FIG. 1. Selection of the self-produced targets for subject CD. (a) The synthetic sounds (empty circles) and the mean locations of American English vowels from Hillenbrand et al. (1995). (b) The imitations of the synthetic sounds (tips of the lines) and the 1-sd ellipses for the prototypical vowels of the subject (hatched ellipses). (c) The subject’s imitations (empty circles), their convex hull with its four corner points (solid line; see text), and the 5×3 vowel grid (dotted lines). (d) The 45 imitations chosen to be the targets.
Each subject was asked to read heed, hid, head, had, hud, hod, hoed, hood, and who’d five times in a list context and another five times in a sentence context. These readings provided 10 tokens of the subject’s natural productions of each of the 9 monophthongal vowels in American English.

2. Imitation of the synthetic stimuli [Fig. 1(b)]

The 100 stimuli were randomized and divided into 2 blocks of 50 stimuli each. The same randomized block and presentation order was used for all subjects. Because of the difficulty of imitating the synthetic stimuli, each stimulus was presented and imitated twice consecutively. This allowed subjects two chances to imitate each synthetic stimulus; only the second imitation was used subsequently. After each stimulus presentation, the subject had 4 s to initiate his imitation. The recording program registered the end of an imitation using an amplitude threshold, waited one second, and then presented the target for the next trial. Subjects were encouraged to produce all their imitations at the same level of subjective loudness; if an error occurred during the recording, the affected targets were presented again at the end of the block. Prior to the imitation session, the subjects were familiarized with the synthetic stimuli and the procedure in a brief training session in which they imitated 15 randomly chosen synthetic stimuli.

3. Selection of the self-produced targets

The F1/F2 “vowel quadrilateral” of each speaker was approximated by the convex hull of the 100 imitations. Four points on the hull perimeter closest to the speaker’s heed, who’d, had, and hod ellipses were chosen as the four “corners” of the quadrilateral. The top and bottom lines of the quadrilateral were each divided into three equal segments and the front and back lines were divided into five equal segments, resulting in 15 cells arranged in a $5 \times 3$ grid [Fig. 1(c); cells are denoted by row#-col#]. The utility of this division is that the $5 \times 3$ grid is qualitatively “normalized” to the vowel space of each subject. For example, all the speakers have [a]-like tokens in cell 3-2 and [ʌ]-like tokens in cell 4-3.

For each subject, three productions were selected from each cell of the vowel grid, for a total of 45 targets [Fig. 1(d)]. The targets were chosen so that they covered a wide range of phonetic qualities. In addition, care was taken to exclude productions that were breathy or creaky, or had sharp variations in pitch and loudness (however, the durations of the targets were not controlled). If a cell did not have at least three productions [e.g., cell 2-1 in Fig. 1(d)], then the closest productions from adjacent cells were assigned to it. Finally, the targets were numbered from 1 to 45, with the three targets assigned to each cell ordered by decreasing F2. For example, the targets belonging to cells 1-1 and 2-1 were numbered [1,2,3] and [10,11,12], respectively.

4. Imitations of the self-produced targets

The presentation list for the imitations of the self-produced targets contained 10 instances of each of the 45 unique targets. The sequence of 450 stimuli was randomized and divided into 9 blocks of 50 each. The protocol for each imitation trial was same as for the synthetic stimuli, and the same randomized block and presentation order was used for all subjects. The use of the same presentation order for all subjects was an approximate control for vowel context effects. Targets with the same number have grossly similar phonetic quality for all speakers, e.g., the target sequence [3,26,43] denotes a [high-front, mid-back, low-back] vowel sequence. This control is not very strict, of course, but it was judged better than allowing subjects to have arbitrarily different vowel contexts. Prior to the imitation session, there was a training session with 15 randomly chosen self-produced targets.

D. Formant analysis

The formants for each imitation were estimated using a customized LPC analysis tool. A 256-point (25.6 ms) analysis frame was slid along the signal in steps of 64 points (6.4 ms). The analysis frame was Hamming-windowed, pre-emphasized at 100%, and submitted to LPC analysis. The optimal filter order was determined separately for each subject (Vallabha and Tuller, 2002). Next, the spectrum of the LPC filter was computed with a 512-point FFT and the locations of its peaks were estimated using three-point parabolic interpolation. The formant tracks were overlaid on a spectrogram to ensure that the LPC peaks accurately captured the spectral structure of the signal; if there were discrepancies, the filter order was adjusted. Finally, a segment containing at least five pitch pulses was selected from the least-varying portion of the production (i.e., with relatively flat F1 and F2 trajectories), and the formant estimates were averaged over the corresponding analysis frames. If the imitations were diphthongized, then the selection was made from the portion of the diphthong that most closely matched the quality of the target. Fortunately, most of the imitations were monophthongs with unambiguous steady states.

III. MODELS

The results from Repp and Williams (1987) suggest that we can expect systematic bias and noise in the imitations. In order to draw interesting conclusions about the underlying mechanisms, it is necessary to formulate plausible null hypotheses. The two models below are a first attempt at such a formulation, and they predict that any patterns observed in vowel imitation are due to noise in tongue positioning during vowel production and/or vowel perception.

A. Articulatory model

We simulated the effect of random articulatory perturbations using the ASY articulatory synthesizer (Rubin et al., 1981). The vocal tract (VT) configurations for six American English vowels [i e æ a u] were obtained, and the AUTO command in ASY was used to compute their formants. In order to increase coverage of the formant space, a raised version of [a] was created and the VT configuration for [a] was made more backed (these two vowels are denoted [a’]).
seven key vowels up a VT configuration were linearly interpolated between the artificial vocal tract. The 10 ASY parameters which make approximately mark the perimeter of the "vowel space" of center. The sd of the noise was 1 mm for both coordinates, serve as "targets," yielding a total of 90 targets.

Finally, six vowels were chosen from each of the 15 cells to serving the same method as with the speakers’ productions. Finally, six vowels were chosen from each of the 15 cells to serve as "targets," yielding a total of 90 targets [Fig. 2(b)].

Each target was "imitated" (perturbed) 10 times. For each perturbation, zero-mean Gaussian noise was independently added to the x and y coordinates of the tongue body center. The sd of the noise was 1 mm for both coordinates, which is a conservative estimate based on an x-ray microbeam study of the production of American English vowels (see Table IV of Beckman et al., 1995). The use of the same noise sd for both coordinates is admittedly unrealistic. However, unequal radii are not meaningful unless the noise el-

and [_y], respectively; the latter symbol represents the close-mid back unrounded vowel). The vowels [i e æ ə a’ y u] approximately mark the perimeter of the "vowel space" of the artificial vocal tract. The 10 ASY parameters which make up a VT configuration were linearly interpolated between the seven key vowels [Fig. 2(a)], resulting in 196 vocalic sounds. A vowel grid was constructed with these sounds using the same method as with the speakers’ productions. Finally, six vowels were chosen from each of the 15 cells to serve as "targets," yielding a total of 90 targets [Fig. 2(b)].

Each target was "imitated" (perturbed) 10 times. For each perturbation, zero-mean Gaussian noise was independently added to the x and y coordinates of the tongue body center. The sd of the noise was 1 mm for both coordinates, which is a conservative estimate based on an x-ray microbeam study of the production of American English vowels (see Table IV of Beckman et al., 1995). The use of the same noise sd for both coordinates is admittedly unrealistic. However, unequal radii are not meaningful unless the noise el-

and [y], respectively; the latter symbol represents the close-mid back unrounded vowel). The vowels [i e æ ə a’ y u] approximately mark the perimeter of the "vowel space" of the artificial vocal tract. The 10 ASY parameters which make up a VT configuration were linearly interpolated between the seven key vowels [Fig. 2(a)], resulting in 196 vocalic sounds. A vowel grid was constructed with these sounds using the same method as with the speakers' productions. Finally, six vowels were chosen from each of the 15 cells to serve as "targets," yielding a total of 90 targets [Fig. 2(b)].

Each target was "imitated" (perturbed) 10 times. For each perturbation, zero-mean Gaussian noise was independently added to the x and y coordinates of the tongue body center. The sd of the noise was 1 mm for both coordinates, which is a conservative estimate based on an x-ray microbeam study of the production of American English vowels (see Table IV of Beckman et al., 1995). The use of the same noise sd for both coordinates is admittedly unrealistic. However, unequal radii are not meaningful unless the noise el-

and [y], respectively; the latter symbol represents the close-mid back unrounded vowel). The vowels [i e æ ə a’ y u] approximately mark the perimeter of the "vowel space" of the artificial vocal tract. The 10 ASY parameters which make up a VT configuration were linearly interpolated between the seven key vowels [Fig. 2(a)], resulting in 196 vocalic sounds. A vowel grid was constructed with these sounds using the same method as with the speakers' productions. Finally, six vowels were chosen from each of the 15 cells to serve as "targets," yielding a total of 90 targets [Fig. 2(b)].

Each target was "imitated" (perturbed) 10 times. For each perturbation, zero-mean Gaussian noise was independently added to the x and y coordinates of the tongue body center. The sd of the noise was 1 mm for both coordinates, which is a conservative estimate based on an x-ray microbeam study of the production of American English vowels (see Table IV of Beckman et al., 1995). The use of the same noise sd for both coordinates is admittedly unrealistic. However, unequal radii are not meaningful unless the noise el-

and [y], respectively; the latter symbol represents the close-mid back unrounded vowel). The vowels [i e æ ə a’ y u] approximately mark the perimeter of the "vowel space" of the artificial vocal tract. The 10 ASY parameters which make up a VT configuration were linearly interpolated between the seven key vowels [Fig. 2(a)], resulting in 196 vocalic sounds. A vowel grid was constructed with these sounds using the same method as with the speakers' productions. Finally, six vowels were chosen from each of the 15 cells to serve as "targets," yielding a total of 90 targets [Fig. 2(b)].

Each target was "imitated" (perturbed) 10 times. For each perturbation, zero-mean Gaussian noise was independently added to the x and y coordinates of the tongue body center. The sd of the noise was 1 mm for both coordinates, which is a conservative estimate based on an x-ray microbeam study of the production of American English vowels (see Table IV of Beckman et al., 1995). The use of the same noise sd for both coordinates is admittedly unrealistic. However, unequal radii are not meaningful unless the noise el-

and [y], respectively; the latter symbol represents the close-mid back unrounded vowel). The vowels [i e æ ə a’ y u] approximately mark the perimeter of the "vowel space" of the artificial vocal tract. The 10 ASY parameters which make up a VT configuration were linearly interpolated between the seven key vowels [Fig. 2(a)], resulting in 196 vocalic sounds. A vowel grid was constructed with these sounds using the same method as with the speakers' productions. Finally, six vowels were chosen from each of the 15 cells to serve as "targets," yielding a total of 90 targets [Fig. 2(b)].

Each target was "imitated" (perturbed) 10 times. For each perturbation, zero-mean Gaussian noise was independently added to the x and y coordinates of the tongue body center. The sd of the noise was 1 mm for both coordinates, which is a conservative estimate based on an x-ray microbeam study of the production of American English vowels (see Table IV of Beckman et al., 1995). The use of the same noise sd for both coordinates is admittedly unrealistic. However, unequal radii are not meaningful unless the noise el-

and [y], respectively; the latter symbol represents the close-mid back unrounded vowel). The vowels [i e æ ə a’ y u] approximately mark the perimeter of the "vowel space" of the artificial vocal tract. The 10 ASY parameters which make up a VT configuration were linearly interpolated between the seven key vowels [Fig. 2(a)], resulting in 196 vocalic sounds. A vowel grid was constructed with these sounds using the same method as with the speakers' productions. Finally, six vowels were chosen from each of the 15 cells to serve as "targets," yielding a total of 90 targets [Fig. 2(b)].

Each target was "imitated" (perturbed) 10 times. For each perturbation, zero-mean Gaussian noise was independently added to the x and y coordinates of the tongue body center. The sd of the noise was 1 mm for both coordinates, which is a conservative estimate based on an x-ray microbeam study of the production of American English vowels (see Table IV of Beckman et al., 1995). The use of the same noise sd for both coordinates is admittedly unrealistic. However, unequal radii are not meaningful unless the noise el-
configuration to a speech waveform, random perturbations of the tongue can potentially lead to systematic biases in the formants. Hence, the articulatory model can be taken as a nontrivial hypothesis about imitation bias patterns and imitation variability. The perceptual model incorporates assumptions about the Bark transformation and formant discrimination limen, so it too can be taken as a nontrivial hypothesis about imitation variability. However, it is incapable of generating any pattern of bias. This incapacity is inherent in the definition of the model and is not a “prediction,” so bias simply does not come with the scope of the model.6

IV. RESULTS

A. Qualitative results

In presenting the results, we concentrate on F1 and F2 patterns since only the F1 and F2 locations of the targets were controlled. When examining the results, it is important to keep in mind that we do not know whether the subjects perceived any differences between the targets and their imitations as they were producing them. One workaround would be to have the subject evaluate the targets and imitations in a post-experiment discrimination task, but this is also problematic since natural productions vary along dimensions other than phonetic quality.

Figures 3 and 4 summarize the imitation behavior of the subjects. Each arrow will henceforth be referred to as a “bias vector.” The principal component ellipses7 show the variation around the corresponding means, and are shown separately in order to make the overall bias pattern more salient. There are five points to note in the figures:

1. All three subjects exhibit a prominent and systematic bias over their entire vowel space. Even when a bias vector is not significant, its direction is usually consistent with adjacent bias vectors;
2. The bias vectors do not seem to be influenced by the nearest prototypes. For example, CD’s imitations of high-back targets (cells 1-3 and 2-3) ignore the [u], [u]...
and [o] prototypes, while DR’s imitations of high front targets (cell 2-1) ignore the [I] prototype;

(3) the biases sometimes appear to centralize (e.g., the high vowels for CD and DC), but in several cases they are directed away from the center (e.g., the bias patterns in CD’s cell 4-3, DR’s cell 4-2, and FC’s cell 3-3 and 5-3). Moreover, the “Day 3” bias pattern for CD (Fig. 7) clearly shows a general lowering rather than a neutralization;

(4) some patterns are interesting by their absence: Low-back vowels are lowered or raised but rarely move directly to the center, and in no case does a mid-back vowel move towards the high-back region (mid-vowels are almost never raised, in general);

(5) the imitations are very noisy. Note especially that the ellipses are 1-sd wide and, therefore, cover only ~40% of the distribution. Moreover, a nonsignificant bias vector does not imply accurate imitation—it usually indicates a lot of variability without any systematic component. Consequently, areas with nonsignificant vectors (e.g., high-central region for CD, low-front for FC) are not necessarily stable;

Figures 4(b) and (c) shows the “imitation” plots of the articulatory and perceptual models overlaid with each subject’s 1-sd principal component ellipses (in order to make the plots less crowded, ellipses are shown for only 76 of the 90 targets). The key point to observe is that the articulatory model’s bias patterns do not match the subjects’ The model’s bias vectors are smaller than the subjects’ and are much less consistent, that is, adjacent targets do not usually move in the same direction. In fact, the ASY model’s bias vectors seem as irregular as the perceptual model’s (recall that by construction, the perceptual model does not exhibit any bias pattern). In addition, the ASY ellipses have a pronounced horizontal orientation, unlike either the perceptual model or the subjects’ productions. The perceptual model’s ellipses are slightly larger for front vowels, but this is only because their sd is defined to be 0.28 barks and the ellipses are being shown in a linear Hertz space.
TABLE I. Statistics of the bias vectors, calculated over the 45 bias vectors (for the subjects) and 90 (for the models). ASY = Articulatory model. PER = Perceptual model.

<table>
<thead>
<tr>
<th></th>
<th>Change in Hertz (mean±sd)</th>
<th>Change in Barks (mean±sd)</th>
<th># significant&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1</td>
<td>F2</td>
<td>F1</td>
</tr>
<tr>
<td>ASY</td>
<td>3±9</td>
<td>3±40</td>
<td>-0.02±0.07</td>
</tr>
<tr>
<td>PER</td>
<td>-2±11</td>
<td>6±23</td>
<td>-0.02±0.09</td>
</tr>
<tr>
<td>CD</td>
<td>13±21</td>
<td>4±73</td>
<td>0.12±0.18</td>
</tr>
<tr>
<td>DR</td>
<td>26±25</td>
<td>-18±64</td>
<td>0.23±0.22</td>
</tr>
<tr>
<td>FC</td>
<td>9±31</td>
<td>-64±66</td>
<td>0.07±0.26</td>
</tr>
</tbody>
</table>

<sup>a</sup>Number of significant bias vectors, using the Hotelling $T^2$ test.

B. Quantitative results

1. Formant bias

Table I shows the F1 and F2 biases averaged over the bias vectors of all the targets. The subjects show significant ensemble effects; specifically, CD and DR tend to lower their imitations while FC tends to retract. This ensemble bias does not, by itself, account for the patterns in Figs. 3 and 4. Subtracting the ensemble bias from all the imitations still leaves a substantial number of significant bias vectors: 22, 18, and 28 for CD, DR and FC, respectively (p<0.01). Gross bias patterns such as the preference for the low-front region are still present after the subtraction (albeit less prominently than before).

In spite of the differences in ensemble bias, there is an underlying similarity among the subjects. This is shown by the bias magnitude, viz. the Euclidean distance from the target to the mean imitation measured in Bark units. Figure 5(a) shows the distribution of the bias magnitude for the models and subjects. Both model distributions are significantly different from the three subject distributions (p<0.001, Bonferroni-corrected Kolmogorov–Smirnov test). None of the other comparisons are significant (with the exception of CD-FC, significant only at $p<0.05$). Figure 5(b) shows that the subjects’ bias magnitudes are larger than the models’ even if the imitation variability is taken into account (the Hotelling $T$ statistic is essentially the bias magnitude in sd units). The pairwise Kolmogorov–Smirnov tests confirm that the only significant differences are between the model and subject distributions ($p<0.001$).

One concern with the bias vector is that it can misrepresent the overall directionality of the imitations. Since it weights the directions of the individual vectors by their magnitudes, it cannot determine whether the imitations occur in one preferred direction, or in all directions with those in one direction being much larger. This concern is addressed by the circular variance, which only uses the unit vectors and ranges from 0 (the ten unit vectors all have the same direction) to 1 (the vectors are maximally dispersed). Figure 5(c) shows the distribution of this measure, and illustrates that the subjects’ imitations usually have a preferred direction with respect to the target whereas the models do not. Again, the only significant differences are between the model and subject distributions.

Finally, there is an intriguing negative correlation between the circular variance and the mean magnitude of the individual vectors ($r=-0.43$, $-0.52$, and $-0.53$ for CD, DR, and FC, all statistically significant at $p<0.05$). The models’ correlations, on the other hand, are much smaller ($r=-0.13$ and $-0.17$) and neither approached significance. Since circular variance ignores the magnitude and the mean magnitude ignores the directions, there is no a priori relation between these two measures. Hence, these correlations are not artifactual, and reliably indicate that large deviations have a more consistent direction than small deviations.

FIG. 5. Frequency polygons of the (a) Bark magnitudes of the bias vectors, and (b) the corresponding Hotelling T values. The vertical line indicates the 95% critical value. (c) Frequency polygon of the circular variances. ASY = articulatory model. PER = perceptual model. N=90 for ASY and PER, N=45 for the subjects.
There is one other piece of evidence for the independence of F1 and F2. The relation between the direction of the bias vector and the orientation of the major axis of the ellipse is extremely weak for the subjects and models ($p > 0.2$, using a directional correlation coefficient, Jupp and Mardia, 1980). There are locations where the ellipse axis and bias direction are strongly related (e.g., high-back region for CD, mid-central region for FC), but this behavior is not characteristic of imitations over the entire vowel space. In brief, the variability of the subjects is more similar to the formant DL-based perceptual model than to the articulatory model.

3. Other acoustic measures

It is possible that the biases in F1 and F2 were perceptually counteracted by other systematic changes in the acoustics. To evaluate this possibility the durations, fundamental frequencies, and higher formants of the targets and imitations were compared to each other. Briefly, the subjects’ imitations of the (150 ms) synthetic vowels were almost always of greater duration than the target (see Pisoni, 1980, for a similar lengthening effect). In imitating the self-produced targets, CD, DR, and FC showed mean duration changes of 17, 42, and $-13$ ms, respectively, but neither these changes nor the absolute durations were correlated with F1 or F2 changes. Likewise, the F0 of the imitations closely matched the F0 of the self-produced target (CD, DR, and FC had mean F0 differences of 6, 0, and $-1$ Hz), and the F0 differences were not related to F1 or F2.

The higher formants are more interesting. The F3 and F4 of the self-imitations were significantly different from the target, and the bias and variability are similar to that for F1 and F2. For example, the average sd was 0.21, 0.22, and 0.28 barks for F3 (CD, DR, and FC), and 0.23, 0.28, and 0.27 barks for F4. These F3 and F4 changes were not correlated with F1 and F2 changes, reinforcing the conclusion that formants are independently affected by noise at the magnitude of the formant DL. Also, the presence of bias for F3 and F4 suggests that the imitation subtly affects the entire spectrum of the vowel.
V. LONG-TERM EFFECTS

From the data reported above, it is clear that subjects are biased in imitating themselves. However, is the particular pattern of bias exhibited by a subject a transient phenomenon, or is it a relatively durable aspect of their perception and production? An answer to this question would help to pin down the functional basis (if any) of the biases. As a first attempt towards an answer, two of the speakers (CD and FC) were asked to participate in a replication of “Day 2” of the imitation task 14 months after their original participation (this will henceforth be called “Day 3;” DR was not able to participate as he had moved out of state). The experimental protocol, stimulus sequence, and recording and playback apparatus were exactly the same as before.

Figure 7 shows the resulting bias vector patterns. There is a marked difference in CD’s overall pattern (especially in cells 1-3, 3-1, and 5-1) with lowering being much more prominent. The median bias magnitude is also larger (0.44 barks, up from 0.33 barks before). FC also shows changes in the bias directions (e.g., in cells 1-1, 1-2, and 5-4) but the overall pattern is very similar to the original one and there is little change in the median bias magnitude. In contrast, the formant variabilities for Day 3 are quite similar to Day 2 for both subjects (the average F1 and F2 sds for Day 3 are 0.24 and 0.31 barks for CD, and 0.20 and 0.32 barks for FC; cf. Table II). Moreover, the distributions of distance from the mean imitations have Rayleigh parameters of 0.28 for both CD and FC, and they are not significantly different from the original distributions ($p > 0.05$, two-sample K-S test).

These results do not provide a clear answer to the question of bias pattern durability. The durability seems to be subject-specific, possibly depending on the particular imitation strategy adopted by the subject. One interesting consistency between Figs. 3(a), 4(a), and 7 is the significance of the bias vectors—the original and replicated vectors of a target are usually both significant or both insignificant. Note, for example, CD’s low-front and high-central regions and FC’s mid-front region. Even more remarkable are the cases where consistently significant and insignificant targets occur next to each other (e.g., CD cell 3-3, FC cell 3-1). These results suggest that regions of the vowel space have characteristic levels of bias significance and are separated by fairly sharp boundaries, although it is unclear how these are related to phonemes.

VI. DISCUSSION

The main results of the experiments reported here can be summarized as follows:

1. The subjects’ imitations of self-produced vowels exhibit biases similar to those observed by Repp and Williams (1987). The bias magnitudes are significantly larger than those predicted by a model of articulatory noise;
2. the biases do not seem to be influenced by proximal vowel prototypes;
3. the imitations seem to be randomly distributed around the mean imitation, with F1 and F2 having independent Gaussian distributions with sds similar to the vowel formant DL of 0.28 barks. These results closely match the assumptions underlying the perceptual model (moreover, additive Gaussian noise in the formant space seems more plausible as a perceptual than an articulatory phenomenon);
4. the directions and magnitudes of the bias are independent of the formant variability;
5. the subjects vary markedly in their bias pattern and may even change them over time, but the magnitudes of the bias and the variability of the imitations (e.g., F1 and F2 sds) are both remarkably consistent within and across subjects.

These results imply that whatever the origin of the bias, its magnitude seems to be (a) instantiated or “worked out” in a similar manner across different subjects, and (b) overlaid...
by noise that is independent of both the bias and the sound being imitated. The results related to the imitation bias are novel, robust, and unanticipated by existing theories of speech perception and production. The results related to formant variability, on the other hand, fit nicely into existing theories and are conceptually straightforward. We discuss each set of results separately.

A. Formant variability

The distribution of F1 and F2 around the mean imitation indicates that the same vowel stimulus evokes different mimics. Some of these differences may be attributed to vowel context effects (Repp et al., 1979), but it is likely that a large part of the variance stems from noise in the imitation process, and that this noise fluctuates across the vowel space and across different sessions. Pisoni (1980) observed a similar phenomenon—when his subjects imitated the same target in different sessions, the mean imitations were highly correlated across the sessions but the sds were not. Moreover, the noise level in the current experiment seems to be characteristic of imitation—the Hertz sds in Table II are of the same magnitude as those reported by Repp and Williams (1985, 1987) and Pisoni (1980). These similarities are even more remarkable in light of the differences among the experiments in recording and playback setups, analysis methods, experimental protocols, and linguistic backgrounds of the subjects.

In fact, the notion of noisy production and perception is rather uncontroversial. Noise is endemic in vowel production studies (e.g., Beckman et al., 1995) and there is substantial evidence that the same stimulus can generate different percepts. This perceptual variability is clear in vowel experiments that use ratings rather than categorization (e.g., Sawusch and Nusbaum, 1979) and affects even phonetically trained listeners (Laver, 1965). Moreover, the assumption of normally distributed percepts is central to several theories of vowel perception (Macmillan et al., 1988; Chistovich et al., 1966; Maddox et al., 2001; Uchanski and Braida, 1998). What is surprising, however, is that the F1 and F2 of the imitations are uncorrelated with each other (surprising because formants are naturally coupled in production). However, the correlation coefficient only measures linear dependence, so it is still possible that F1 and F2 are nonlinearly related (in particular, a nonmonotonic relation between F1 and F2 would result in a weak correlation). Such relations are difficult to evaluate with small samples such as ours, but characterizing them would be an interesting research issue in its own right.

There are two other surprising results regarding the formant variability. First, the formant variability of the articulatory model is similar to that of the perceptual models, even though the parameters of the two models were independently motivated—the 1 mm perturbation radius of the articulatory model was based on the variability of tongue movement, while the 0.28 bark sd of the perceptual model was based on vowel discrimination. Second, the F1 and F2 standard deviations are commensurate with the formant DL. This confirms speculations made by Kent and Forner (1979), but it also raises an interesting problem. If articulatory precision (σ) is matched to perceptual sensitivity and noise is assumed to be additive Gaussian, then a “round trip” through the system should be subject to noise of at least sqrt(σ² + σ²). The current experiment suggests that the overall noise is still σ (=0.28 barks), which is only possible if perceptual and production noise “balance” each other in some way, or if there are nonlinearities (e.g., phonetic categorization) in the system. In either case, these results suggest an attunement between the production and perceptual systems (cf. Fox, 1982), though it is unclear whether the attunement is due to codevelopment (speakers tacitly learn the precision with which vowels may be produced or perceived) or coevolution (the intrinsic noise levels of the production and perception systems have become matched/balanced over time).

We note in passing that the formant variability of the articulatory model [Fig. 4(b)] shows that high-central vowels are much more variable than vowels in other locations, especially along F2. This result is consistent with predictions from quantal theory (Stevens, 1989) and with experiments in which subjects imitated tokens along the [i-u] continuum (Kent, 1973; Repp and Williams, 1985, 1987).

B. Formant bias

There are three theoretical issues underlying the bias: (1) Why is there bias at all? (2) what is the principle underlying the pattern of directions? and (3) what is the cause of the intersubject differences?

Some putative explanations can be dismissed at the outset. One is that the deviations are introduced into the waveform during recording or playback. Another is that during imitation, a speaker’s self-perception of his production is distorted by bone conduction (Maurer and Landis, 1990), so that he is not really producing what he thinks he is producing. Both these explanations predict a unidirectional bias across the formant space, which conflicts with the diversity of bias directions both within and across subjects. The change in CD’s bias directions between Day 2 and Day 3 also argues against a low-level or artifactual origin of the bias. There is a third strawman explanation, viz. that the movements are simply due to subjects’ mishearing the target (by not paying attention, for example) or misarticulating their imitation. There is some plausibility to this argument, since our subjects reported that they were occasionally aware of a difference between the target and their mimic. However, note that Repp and Williams (1987) reported similar levels of inaccuracy even though their subjects had extensive phonetic training (we also obtained similar results in pilot studies with experienced subjects). Thus, the biases cannot be simply ascribed to a lack of skill.

Another concern is the robustness of the bias and the extent to which one can generalize from the limited number of subjects. In this regard, it should be noted that the key claim of the paper is the statistical significance of the bias and not its precise magnitude or the patterns of bias directions. There is additional evidence for the generality of the bias. First, the subjects in our pilot experiments also showed the bias. In fact, we had a fourth subject (a female speaker) in our experiment who showed bias to a larger degree than the three male subjects. We omitted her data since she had phonetic training and was also acquainted with the purpose.
of the experiment, so we felt this might have affected her imitation behavior. Second, results from prior experiments bear out the fact of the bias (see Vallabha, 2003, Chapters 3 and 4); in these experiments, subjects imitated themselves in a “chained” manner, i.e., the imitation was played back to the subject as the target for the next imitation, and these imitation chains showed systematic drifts. It is a valid concern whether the pattern of bias directions generalizes across subjects; we make no such claims in this paper except to point out some regularities (e.g., it is extremely rare to see a mid-vowel quality shifting to a high-back quality).

A related concern is about the role of dialect differences among the subjects. We had originally assumed that the vowel space would be structured by the gross characteristics of the native language (e.g., the /i-I/ contrast is present in American English but absent in Spanish), and that subtle dialectal variations would be less of an issue. The results of the experiment do not support this view, and leave open the possibility that the intersubject differences in the bias patterns are due to dialectal differences (though the difference between CD’s “Day 2” and “Day 3” does not support this possibility). In general, however, the dialect confound does not affect the larger point that it is surprising to find bias at all.

Below, we consider two different classes of explanation for the bias—production-based and perception-based. This distinction is made only to organize the exposition, and should not be taken as a claim that perceptual and production influences can be cleanly separated (for example, an explanation couched in terms of articulatory synergies can be reworded to refer instead to the perception of the acoustic consequences of the synergies). This caveat is especially pertinent because of the integrated nature of imitation.

1. Production-based explanations

The ASY model of noise only examined the consequences of random noise around each articulatory configuration and showed that such noise does not explain the directionality of the subjects’ imitations. However, ASY models only the gross anatomy and kinematics of the vocal tract and does not (usually) take into account muscles that actually move the articulators or the functional synergies that exist between them. Thus, it is still possible that articulatory noise, shaped by physiological or functional constraints that are omitted from the ASY model, can lead to the kinds of movements seen in the current data.

The issue of physiological noise was addressed by Mooshammer et al. (1999), using a two-dimensional (2D) biomechanical tongue model that included the major tongue muscles and elastic properties of the tissues. In the context of the equilibrium-point motor hypothesis, they added independent signal-related Gaussian noise to the muscle commands and found that the noise does not account for the token-to-token articulatory variability observed with real speakers. Moreover, the acoustic variability (i.e., F1/F2 dispersion ellipses) of the Mooshammer et al. model is qualitatively similar to that seen with ASY [Fig. 4(b)]. Thus, physiological noise fails to account for the directionality of the imitations.

Another kind of noise is that shaped by functional constraints. This view posits that vocal tract muscles are controlled in a correlated manner, with the underlying principle being thought of as an abstract “functional variable” (Kelso and Tuller, 1983). Noise in the functional variables would lead to correlated noise among the components, which could cause much more directed variation than observed with uncorrelated noise. For vowels, two commonly used functional variables are the location and degree of the vocal tract constriction (cf. Stevens, 1989; Browman and Goldstein, 1990). If increases and decreases in constriction degree and location are equally likely, we would expect imitations to be distributed around the target, which is not the case with the current data. It is more plausible that increases and decreases are not equally likely (“functional noise”). If speakers prefer alveolar constrictions and less constricted vocal tracts, for example, we would expect front vowels to centralize and low and back vowels to advance and become slightly more open (Gay et al., 1992; Perkel and Nelson, 1985). This notion is plausible and intriguing, but it is hampered by a lack of evidence. In addition, we would need to fit a different set of biases for each subject in order to explain the data. These objections are pragmatic and not conceptual, however, so the functional noise explanation cannot be ruled out completely.

Finally, observe that quantal theory (QT; Stevens, 1989) is also an unlikely explanation of the observed formant movements. QT tries to explain how point vowels such as [i] and [u] are stable and therefore prevalent in vowel systems, but the current data show subjects moving away from the [i] and [u] regions. Further, QT predicts that the vowel space is acoustically more stable near the point vowels [i], [a], and [u], but the subjects’ variability ellipses do not show any such contrast between point and nonpoint vowel regions (Pisoni, 1980; but see Diehl, 1989).

2. Perception-based explanations

Imitation directionality has traditionally been explained as assimilation caused by a categorical phonemic code. Similarly, the perceptual magnet effect (Iverson and Kuhl, 1995) predicts perceptual assimilation towards the phonemic prototypes. Alternatively, if perception and production are seen as sharing a common control space (e.g., motor theory), then key locations of this space may function as attractors. Yet all these explanations fall short because in the two-dimensional space it is clear that movements are not always influenced by the nearest phoneme (and in any case, the bias direction patterns are too different across the subjects).

It is intriguing to consider the converse situation, namely, that the imitations are moving away from the prototypes and towards “perceptual anchors” located at the category boundaries (Macmillan et al., 1988). Figure 3 shows some evidence of such movement—for example, the “attractors” in CD’s low-back region and DR’s low-front region seem to be at the edges of the [a] and [æ] categories, respectively. It is also striking that the regions between the prototypes are so well-populated (the data from Repp and Williams, 1987, show a similar pattern). This proposal of “boundary attraction” is related to a theory advanced by Cowan and Morse (1986) to explain stimulus order effects. They proposed that a percept is initially a tightly bounded...
region in the perceptual space and that as it decays, it expands and stretches towards the “neutral point” [6]. Repp and Crowder (1990) tested this hypothesis and found that multiple neutral points were needed to explain their results. Intriguingly, they observed that “vowels held in memory were assimilated toward some standard(s) located between prototypes” (p. 2088). This suggests that the neutral (or stable) points may arise from an interaction between perceptual representations and category boundaries, and that the imitation biases are shaped by the locations of these points. While this theory is plausible, it should be kept in mind that imitations are unaffected by response latency (Repp and Williams, 1987). Representational decay may explain the bias directions, but some other factor is needed to explain the bias magnitudes.

There is one other potential explanation for the bias. In a recent study, Dissard and Darwin (2000) played synthetic one and two-formant sounds to subjects, and asked them to match the target by adjusting the formant frequency of a comparison sound. Interestingly, the matched sound deviated systematically from the target when the two sounds had different FOs. Dissard and Darwin suggested that this occurs because listeners’ estimate of the formant location is biased towards the closest FO harmonic. However, this explanation predicts that because there are fewer harmonics in the range of F1 frequencies, the biases should be more prominent across F1 than F2. This is not the case with the data (see Table I). Moreover, the FO differences between the targets and imitations are quite small (less than 5 Hz).

VII. CONCLUSIONS

The current experiment has shown several robust effects in vowel imitation, three of which are particularly significant—formant standard deviations commensurate with formant difference limens, systematic inaccuracy of the imitations, and independence of variability and accuracy of the imitations. These effects confirm patterns found in previous studies and demonstrate the utility of using a grid of stimuli rather than a single continuum.

The subjects and the models have very similar degrees of variability (~0.28 bark Euclidean distance), indicating a convergence shaped by physiology, development and/or linguistic environment. The analyses also demonstrate that it is not sufficient to report statistics (such as standard deviations or histograms) for single formants, as was the procedure in prior vowel imitation studies; some of the statistical patterns are evident only by treating the perceptions as two-dimensional points. The results also reinforce the view that bark distance is an approximate measure of perceptual distance.

The biases of the imitations are also instructive. They cannot be explained by random articulatory noise, nor are they consistent with theories of categorical phonemic codes. There are two explanations (functional noise and perceptual anchors) that cannot be rejected on the current evidence, but they are both are quite tentative and require further elaboration. In this context, two limitations of the current work must be noted. The first is the confound between individual differences and dialect differences. Controlling for dialect background is necessary to clarify the linguistic relevance of the bias patterns. The second limitation concerns the role of isolated steady-state vowels. Such sounds are unusual in any language, and the imitation bias may partly be the result of the speakers’ perception-production system contorting itself to an unusual task. More natural target stimuli, such as vowels embedded in a syllabic context, may yield more consistent results both within and across subjects.

ACKNOWLEDGMENT

This research was supported by NIMH Predoctoral Training Grant No. MH19116 and NIMH Grant No. MH42900.

1Subjects read two sets of sentences, one repeated thrice, the other twice. The thrice-repeated sentences were: “I had hoed my field and did not heed the hood. I hid from Hod, who’d been the Head at Hud.” The twice-repeated sentences were: “Who’d hoed the field? I had the hood on my head. Hod hid in the hood, and did not heed Hud.”
2CD and FC’s imitations of the synthetic stimuli were available from previous pilot experiments and were reused (these were collected using the same protocol described here). The interval between “Day 1” and “Day 2” was 20 months (CD), 26 months (FC) and 2 days (DR). A point of concern is whether CD and FC’s imitations were affected by the delay. Two points mitigate this concern. (1) we conducted several pilots where the delay between Day 1 and Day 2 was quite short, and all the pilot subjects exhibited bias and variability patterns that were qualitatively similar to CD and FC. (2) CD and FC’s performance in the pilots, recorded shortly after their “Day 1,” is also similar to their performance reported here.
3The parameters were obtained from the ASY web site: http://www.haskins.yale.edu/Haskins/MISC/ASY/DYNAMIC/samples.html
4In ASY, the location of the tongue body center (TBC) is defined with respect to an arbitrary origin near the temporomandibular joint (the parameters are CL and CA, for length and angle, respectively). The tongue body rides on the jaw, so the true angle of the TBC is the sum of the jaw angle (JA) and CA. Hence, $x=\text{CL}\cdot \cos(JA+CA)$ and $y=\text{CL}\cdot \sin(JA+CA)$. CL is units of mermels (1 mermel=1/112 cm=0.09 mm). Once the noise was added, $x$ and $y$ were converted back to polar coordinates: $(JA+CA)_{\text{new}}=\text{arctan}(y/x)$, and $\text{CL}_{\text{new}}=\text{sqrt}(x^2+y^2)$. We wanted to perturb only the TBC and not the jaw, so $\text{CL}_{\text{new}}=(JA+CA)_{\text{new}}-\text{JA}$. For a related approach, see Goldstein (1983).
5It would have been preferable to use DLs based on correlated patterns of formant changes rather than single-formant change. In one of the few such studies (Hawks, 1994), the results were established using trained listeners under minimal uncertainty, so the DLs were extremely small (~0.15 barks). We used the larger DL (0.28 barks) since it was based on more realistic methods (“ordinary listening conditions”).
6Ideally, the perceptual model should have incorporated the warping of the perceptual space due to the native vowel categories (e.g., Kawey-Pont and Atal, 1989). Such a warping may produce imitation bias, and therefore increase the scope of the perceptual model. However, we are not aware of any models that attempt to link acoustics to subcategorical vowel qualities over the entire vowel space, and developing and validating such a model is beyond the scope of this paper.
7If we assume that the ten F1xF2 points come from a bivariate Gaussian distribution, then the principal components are the axes of the distribution, and the ellipses denote a radius of one standard deviation along the principal components. Equivalently, the ellipses are the equal-likelihood contours of the Gaussian distribution (likelihood=0.9925).
8The circular variance is computed as follows (Fisher, 1993). Each set of 10 imitations yields 10 unit vectors with respect to the target. The magnitude of their vector sum divided by the number of vectors gives the mean resultant length $R$. The circular variance is defined as $1-R$ and ranges from 0 (all vectors have the same direction) to 1 (the directions of the vectors cancel out each other). If there were no bias, then we would expect the different unit vectors to have inconsistent directions, and therefore the circular variance would be close to 1.

Figure 6 does not show the distribution for the perceptual model since it is Rayleigh-distributed by definition. In general, $x_1, x_2, \ldots, x_K \sim N(0,\sigma)\Rightarrow x - N(0,\sigma\sqrt{(K-1)/K})$. For the perceptual model, $K$

G. K. Vallabha and B. Tuller: Perceptuomotor bias in vowel imitation

1197


Vallabha, G. K. (2003), Perceptuomotor biases in vowel imitation, Unpublished Ph.D. Dissertation (Florida Atlantic University, Boca Raton, FL).

