Perception & Psychophysics
2004, 66 (3), 422-429

Imitation in shadowing words

KEVIN SHOCKLEY
University of Cincinnati, Cincinnati, Ohio
Haskins Laboratories, New Haven, Connecticut
and University of Connecticut, Storrs, Connecticut

and

LAURA SABADINI and CAROL A. FOWLER
Haskins Laboratories, New Haven, Connecticut
and University of Connecticut, Storrs, Connecticut

Imitation of shadowed words was evaluated using Goldinger's (1998) AXB paradigm. The first experiment was a replication of Goldinger's experiments with different tokens. Experiment 1's AXB tests showed that shadowed words were judged to be better imitations of target words than were baseline (read) counterparts more often than chance (.50). Order of presentation of baseline and shadowed words in the AXB test also significantly influenced judgments. Degree of prior exposure to token words did not significantly influence judgments of imitation. Experiment 2 employed modified target tokens with extended voice onset times (VOTs). In addition to AXB tests, VOTs of response tokens were compared across baseline and shadowing conditions. The AXB tests revealed shadowed words to be better imitations of target tokens than baseline, without an influence of AXB presentation order. Differences between baseline and shadowing VOTs were greater when VOTs were extended. The implications of spontaneous imitation in nonsocial settings are considered.

Acoustic signals corresponding to identical nominal utterances may vary dramatically. These variations occur across speakers, since vocal tracts differ, most notably because of differences in dialect, gender, and age and, of course, according to individual differences. Additional variability is present within speakers, due to changes in phonetic context, prosody, and speaking rate.

Some of these differences allow speakers to sound more or less like one another, and research findings suggest that speakers/listeners may exploit this possibility in their social interactions. Speakers "accommodate" (converge) in their accents, speaking rates, rates of pausing, and vocal intensity in cooperative conversations (see Giles, Coupland, & Coupland, 1991, for a review). Moreover, interlocutors do not limit their convergences to speech. They are said to move in synchrony with each other's speech rhythms (Condon, 1976) and to match one another's postures (LaFrance, 1982; Shockley, Santana, & Fowler, 2003).

These tendencies may have a social, or cooperative, motivation. However, they appear to rest on a more fundamental disposition to imitate whether or not the setting is a social one. Indeed, imitation is pervasive among humans. Imitation of facial gestures (e.g., tongue protrusion) is observed in neonates (e.g., Meltzoff & Moore, 1983). Moreover, by 12 weeks of age, infants imitate vocalic sounds (Kuhl & Meltzoff, 1996). Meltzoff and Moore (1997) have suggested that imitation is "a mechanism for intergenerational transmission of acquired characteristics" (p. 179). Whether it is or not, the tendency to imitate persists into adulthood. For example, adults imitate facial expressions (McHugo, Lanzetta, Sullivan, Masters, & Englis, 1985) and speech (Goldinger, 1998).

We propose to look for the occurrence and the nature of imitative responses in perceived acoustic speech signals or in perceived speech gestures by using a shadowing procedure. Early research by Chistovich, Klass, and Kuz'min (1962), followed up by that of Porter and Castellanos (1980) and Porter and Lubker (1980), has shown that listeners can shadow spoken targets remarkably quickly in choice, as well as simple, response time paradigms. Research in general shows that choice response times are longer by 100–150 msec than are simple response times (see also Luce, 1986). In a choice response time task, participants make distinct responses to distinct stimuli; accordingly, a stimulus having been identified, there remains the choice as to what response should be made. For example, participants might be instructed to hit one response button to a green light and a different button to a red light. In a simple response time task, they hit the same response button regardless of the

The research was supported by NIDCD Grant DC-03782 to Haskins Laboratories. Correspondence concerning this article should be addressed to K. Shockley, Department of Psychology, University of Cincinnati, ML 0376, 429 Dyer Hall, Cincinnati, OH 45221-0376 (e-mail: kevin.shockley@uc.edu).

Note—This article was accepted by the previous editorial team, headed by Neil Macmillan.

color of the light. It is not surprising that choice response times tend to be longer than simple response times. What is surprising is that the response time difference may nearly disappear in tasks in which simple and choice stimuli and responses are spoken.

In the research by Porter and Lubker (1980), participants began producing the vowel /a/ as soon as a model began to produce it; when the model shifted to a production of /i/, /ae/, or /o/, the simple response was to produce /o/; the choice task was to produce /i/, /ae/, or /o/, respectively. Under these conditions, response times in the two tasks were very short (169 msec when the simple response was /o/ and 181 msec when the choice response was /o/) and were in the vicinity of simple response times in other tasks.

Porter and colleagues and Chistovich and colleagues suggested that the findings are most consistent with a theory of speech perception in which listeners perceive the articulations of the speaker (e.g., Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). In that case, the task provides an instance of extreme stimulus–response compatibility. In the choice task, for listeners who perceive speech gestures, the stimulus in effect provides instructions for the response. That is, having perceived, say, a high back positioning of the tongue accompanied by lip rounding (for /u/), speakers respond by producing a high back positioning of the tongue accompanied by lip rounding. From the perspective of a theory in which acoustic signals are perceived, the findings are more difficult to explain. Given an acoustic signal, listeners have to decide how to produce a signal that counts as a match by instructing the articulators to move. That is, having perceived a vowel with a low first and second formant, they have to decide how to instruct their articulators so that they produce a vowel with those properties. The task is still a choice response time task and should lead to slower responses than a simple task in which, any stimulus change having been detected, the response is to unleash an invariant set of instructions to the articulators.

A different acoustic account may be possible, however, although we are unaware of any theory in which it is offered. Listeners may identify a phonological category on the basis of acoustic cues, but stored with the category information is information for producing a token of the category. This may be the source of the rapid responding.

Shadowing has been used as a tool not only to explore how quickly an individual can repeat an utterance by another, but also to explore the conditions under which repetitions are imitations of the model speaker. Goldinger (1998) elicited utterances by participants that, in one condition, could not be imitations of a model's speech (the baseline condition) and, in another condition, were repetitions of a model's speech that could, therefore, be imitative. He found that listeners judged the repetitions to be better imitations of the model than were the baseline productions. Moreover, his research (see below) uncovered variables, such as word frequency and the number of times the participant heard a model's word before repeating it, that affected imitative fidelity. He concluded in favor of a theory of speech perception in which listeners store traces of perceived words that retain information about the contexts in which they occur, the voice quality of the speaker, and other idiosyncrasies. The stored traces of a word (exemplars) are activated when the word is spoken, and they coalesce with the spoken word to guide shadowing responses. The shadowed words, therefore, are imitative. Although Goldinger's particular exemplar theory is an acoustic theory, because stored properties of words are features extracted from acoustic signals, a theory similar in spirit might be developed from a gestural perspective. In that case, the phonetic properties extracted from acoustic signals are gestures, not phonetic features.

In the present pair of experiments, we will not attempt to discriminate among the accounts of shadowing speed and imitation that we have reviewed. Rather, our goals are preliminary to that. We hope to replicate and extend Goldinger's (1998) finding that adult's shadowing responses, obtained in nonsocial settings, are imitative. In addition, we hope to begin to assess what is imitated. Specifically, we will ask whether voice onset time (VOT) variability—a subphonemic, phonetic property of words—is imitated.

EXPERIMENT 1

This experiment was a replication of recent experiments by Goldinger (1998) with new testing materials. Goldinger reported that words produced as shadowing responses to target words are identified by listeners as imitations of the target words more often than are productions of the same words that were not produced as shadowing responses. Because Goldinger's findings were strongest for low-frequency words (100 occurrences or less in Kučera & Francis, 1967), all the words for this experiment have frequencies of less than 75 per million (Kučera & Francis, 1967). Goldinger also reported that the number of prior exposures to the target word (2, 6, or 12 exposures) systematically increased listeners' tendency to select the shadowed words as imitations. Accordingly, we also manipulated the number of exposures to target words prior to shadowing. In anticipation of Experiment 2, we used only words beginning with a voiceless stop consonant.

Following Goldinger (1998), we used an AXB test to obtain listeners' judgments of imitation. In such a test, A, X, and B are all tokens of the same word type, such as *partner*. A and B are produced by the same speaker (shadower), whereas X is produced by a different speaker (model). On every trial, one of the response tokens, A or B, is a baseline token (produced by a shadower before he or she is exposed to the speaker of X), and the other is a shadowed token. The participant is asked to judge which token (A or B) is a better imitation of the target word (X). Because A, X, and B are tokens of the same word and because A and B are spoken by the same individual, the basis for listeners' selections must be any respects in which the shadowed production, in fact, sounds like a better imitation of X than does the baseline production.

Following Goldinger (1998), we expected to find better-than-chance performance by listeners, and we expected performance to be better for words that had been repeated six times in a familiarization phase of the experiment than for unrepeated words. In accordance with Goldinger, we also expected better-than-chance performance even for unrepeated words. This provided our first measure of a (near) on-line imitative tendency. Furthermore, if words repeated 6 times in the familiarization phase were subsequently imitated with more fidelity (as indexed by higher AXB performance by the listener) than were words not presented during familiarization, we would have replicated Goldinger's findings that memory for exemplar word tokens guides shadowing responses, which, therefore, are imitative of them.

Method

Participants. Two graduate students, 1 male and 1 female, served as model speakers for the target tokens. Eight undergraduates (4 females and 4 males) provided baseline and shadowed response tokens. An additional 16 undergraduates, 8 males and 8 females, performed the AXB test. The graduate students participated on a voluntary basis, and the undergraduates participated in partial fulfillment of a course requirement. All the participants provided informed consent prior to participation.

Materials and Apparatus. Eighty bisyllabic English words with a frequency less than 75 occurrences per million (Kučera & Francis, 1967) and beginning with the voiceless stop consonants /p/, /t/, and /k/ were used. All the words and their frequencies are listed in Appendix A.

Procedure. Two speakers, one female and one male, served as model speakers of the 80 token words. The words were presented, 1 at a time, on a computer screen every 2 sec, using PsyScope (Cohen, MacWhinney, Flatt, & Provost, 1993). The speakers were instructed to "identify the word you see by speaking it into the microphone quickly but clearly."[1] The words were presented in random order. The tokens were initially recorded on cassette tapes, using a microphone and a cassette recorder. These recorded tokens were then digitized using SoundScope (GW Instruments, Somerville, MA) for use as target tokens in the experiment proper.

The experiment proper proceeded in four phases. In the first phase, baseline productions of the same 80 words that the original speakers had produced were collected from 8 participants (4 females and 4 males). These were collected in the same way as that in which the target words had been collected from the model speakers. That is, the participants saw each of the 80 words presented 1 at a time on a computer screen, and they were instructed to "identify the word you see by speaking it into the microphone quickly but clearly." Responses were recorded on cassette tape and were later digitized. They served as baseline productions.

In the second, familiarization phase, using headphones, the same participants were exposed to zero or six repetitions of the 80 target word tokens (40 words not presented, 40 presented six times) of the original 2 speakers. On each of the 240 trials, the participants heard a word spoken by 1 of the original speakers. For a given participant, when words were repeated, repetitions were always by the same speaker. The female participants were exposed to tokens from the female model speaker, whereas the male participants were exposed to tokens from the male model speaker. The 40 word sets were counterbalanced across all participants, so that half of the participants received one set of 40 words six times and the other half of the participants received the remaining set of 40 words six times. The participants were instructed to "identify the word you hear by writing it next to the appropriate trial number on the response sheet." In the third phase of the experiment, all 80 tokens of the corresponding model speaker were presented to the participants, using headphones. The participants were instructed to "identify the word you hear by speaking it into the microphone quickly but clearly." These shadowed responses, like the baseline responses, were recorded on audio tape and were later digitized.

The fourth phase of the experiment presented an AXB test to a new set of participants. From each set of baseline words, target words, and shadowed words of Phases 1 and 3, separate AXB tests were devised from the responses of each participant. That is, eight different AXB tests were constructed. Each of 16 different listeners completed one of the eight distinct AXB tests. Tests consisted of 160 trials in which each of the 80 target words appeared as X (the model the speaker shadowed) twice, once with A as the baseline production and B as the shadowed word and once with A as the shadowed word and B as the baseline. Token words were presented over headphones, using PsyScope. The listeners were asked to indicate which of A or B was the better imitation of X. Responses were collected using a PsyScope button box, consisting of three adjacent buttons. The left button was pressed to indicate that the first item (A) was a better imitation of the target token (X), whereas the last button was pressed to indicate that the last item (B) was a better imitation of X. The middle button was inactive.

Results

The AXB test revealed that the mean percentage of shadowed tokens reported to be better imitations of target tokens (X) was 56% for zero prior presentations and 58% for six prior presentations. The percentages of shadowed tokens reported to be better imitations of target tokens (X) were compared with chance (50%), using t tests. For zero and six prior presentations of the target word, the shadowed tokens were reported to be better imitations of the target tokens (X) [$t(15) = 2.57, p < .05$, and $t(15) = 3.01, p < .005$, respectively]. (In this and all subsequent statistical tests, we set alpha to .05.)

We also evaluated the role of the position of the baseline and the shadowed words in the AXB test. Mean values for all configurations can be seen in Table 1. When the baseline tokens were presented first (A) and the shadowed tokens were presented last (B), the shadowed tokens were reported to be better imitations of the target tokens (X) [$t(15) = 3.45, p < .005$, for zero prior exposures, and $t(15) = 3.77, p < .001$, for six prior exposures]. However, when shadowed tokens were presented first (A) and the baseline tokens were presented last (B), the shadowed tokens were not identified as better imitations of the target tokens more often than chance [$t(15) = 1.32$, $p = .10$, and $t(15) = 0.25, p = .40$, for zero and six repetitions, respectively]. The influence of the number of prior presentations was evaluated by comparing zero

Table 1
Mean Percentages of Shadowed Tokens Reported as Better Imitation of Target Token (X) and Standard Deviations

	Presented First in AXB Test							
	Baseline Tokens				Shadowed Tokens			
	Zero Prior		Six Prior		Zero Prior		Six Prior	
Experiment	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	61	12	65	13	51	12	51	12
2	61	13	58	15	63	13	60	14

versus six prior presentations [$t(15) = 0.36, p = .73$, and $t(15) = 0.430, p = .68$, respectively].

Discussion

As was expected, shadowed words were reported as better imitations of target words, as compared with baseline words, replicating Goldinger's (1998) finding. Unlike Goldinger, we did not find evidence that the number of prior exposures to the token words (zero vs. six) significantly influenced the reports of better imitation. Goldinger used a greater variety of repetitions (between 0 and 12); however, his figures suggest an effect of zero versus six repetitions.

There are differences in our procedures and those of Goldinger (1998) that may underlie the difference in outcome. Relevant ones may be that each of our listeners heard just 1 model speaker, whereas Goldinger's listeners heard 10, and that our words were more uniform in their phonetic properties than were Goldinger's, in that all of our words began with voiceless stops. In an exemplar account of the memory traces established in our experiment, there were many more that shared voicing features and initial consonant features with words repeated zero or six times than in Goldinger's experiments. In the exemplar theory that Goldinger tested (Hintzman's, 1986, MINERVA), a spoken word generates an echo composed of words in memory that share features with the input word. Words in memory will contribute to the echo to an extent that depends on the number of features they share with the input word. In Goldinger's account, the echo guides production of a shadowing response to a word. Because of the greater uniformity of voicing and initial consonant in our experiment, more words contributed more influentially to the echo than in Goldinger's research, and this may have washed out the effect of repeating a given word six times. Further research will be needed to determine the conditions under which an effect of repetition can be obtained.

We found that the order of token presentation in the AXB test (baseline tokens presented first vs. shadowed tokens presented first) had an impact on performance, so that the shadowed token was reported to be a better imitation of the target token (X) only when the shadowed token was presented last (B) in the AXB test. This hints at a memory interpretation. Listeners are more accurate when the critical token occurs most recently. Goldinger (1998) did not report results separately by order. However, Goldinger (personal communication, July 29, 2002) has reexamined his data and reports a marginal (zero repetitions) and nonsignificant (six repetitions) trend in the same direction as that for our findings in the present experiment. To complicate matters further, however, in Experiment 2, below, we did not replicate this outcome. Because it is not central to the aims of this research, we do not consider further the reasons why it may have occurred in Experiment 1.

EXPERIMENT 2

A limitation of Goldinger's (1998) AXB procedure is that it does not provide the opportunity to evaluate what is imitated. To determine whether variables other than surface durations of words or their intonation contours are imitated, the present experiment replicated Experiment 1, using edited tokens of the target words from Experiment 1. The VOTs of word initial stops were extended. Baseline words were recorded before the participants were exposed to the target words. Thus, if imitation occurred, words should have longer VOTs in the shadowing condition than in the baseline condition. If imitation did not occur, VOTs should be comparable in the two conditions.

Method

Participants. In addition to the 2 model speakers from Experiment 1, 8 undergraduates (4 females and 4 males) provided baseline and shadowed tokens. None of these 8 undergraduates had participated in Experiment 1. An additional 16 undergraduates, 8 males and 8 females, performed the AXB test. None of these undergraduates had participated in Experiment 1. The undergraduates participated in partial fulfillment of a course requirement. All the participants provided informed consent prior to participation.

Materials and Apparatus. The same materials as those in Experiment 1 were used in the present experiment.

Procedure. The same procedure for Phases 1–3 was used as that in Experiment 1, with the following exceptions. The VOTs of the model speakers' target tokens were extended so that they were twice their original duration on average, and the VOTs of the participants' baseline and shadowed tokens were measured by two of the investigators (K.S. and L.S.). Measurement and manipulation of VOT is detailed in Appendix B. The extended VOT words were noticeably breathier than the original productions. Although they did not sound unnatural, given that the VOTs were twice as long as those that our model speakers had produced, they likely would receive lower *goodness* ratings (e.g., Miller & Volaitis, 1989) than would our original tokens.

Results

The AXB test revealed that the mean percentage of shadowed (vs. baseline) tokens reported to be better imitations of target tokens (X) was 62% for zero prior presentations and 59% for six prior presentations. The percentage of shadowed tokens reported to be better imitations of target tokens (X) was compared with chance (50%), using t tests. For zero and six prior presentations of the target word, the shadowed tokens were reported to be better imitations of the target tokens (X) [$t(15) = 4.59, p < .0005$, and $t(15) = 3.06, p < .005$, respectively]. Table 1 shows the percentages of shadowed tokens selected as better imitations, separated by whether the shadowed token was A or B. When the baseline token was presented first (A) and the shadowed token last (B), the shadowed tokens were reported to be better imitations of the target tokens (X) [$t(15) = 3.46, p < .005$, for zero prior exposures, and $t(15) = 2.08, p < .05$, for six prior exposures]. When the shadowed tokens were presented first (A) and the baseline token was presented last (B), the shadowed tokens were reported to be better imitations of the target tokens (X) [$t(15) = 4.15, p < .0005$, for zero prior exposures, and $t(15) = 2.97, p < .005$, for six prior exposures]. There were no significant differences between zero and six prior presentations for any AXB order presentation.

We next asked whether shadowed tokens had longer VOTs than did baseline tokens. This would suggest a

basis for the imitation judgments in the AXB test. The experimental conditions (baseline vs. shadowing) corresponding to each set of tokens were coded with a random number. Two raters were assigned to measure the VOTs of the participants' tokens. The raters were blind to the condition of token sets. To evaluate interrater reliability, both raters measured two common sets of tokens, in addition to their respective assigned sets. The VOTs of two sets of baseline and shadowed tokens were independently measured by the two raters. The correlation between the measured VOTs of the two raters was $r = .92$.

The measured VOTs of the participants in this experiment were compared across the baseline and shadowing conditions. A one-tailed t test revealed significantly longer VOTs for the shadowing condition than for the baseline condition [$t(7) = 4.45, p < .005$]. The mean baseline and shadowed VOTs were 51.8 msec (SD = 12.5) and 64.1 msec (SD = 11.8), respectively. To check the possibility that VOTs may have likewise been extended in the shadowing condition of Experiment 1, the VOTs in that experiment were also measured and compared across the baseline and shadowing conditions. A one-tailed t test did reveal significantly longer VOTs for the shadowing condition than for the baseline condition [$t(7) = 3.16, p < .01$]. The mean VOTs for the baseline and the shadowing condition were 59.5 msec (SD = 7.9) and 65.3 msec (SD = 11.6), respectively.

To compare the results across experiments, difference scores between the baseline and the shadowing VOTs in each experiment were compared using a one-tailed t test. Experiment 2 (extended VOTs for model speakers) showed significantly larger differences in VOTs between the baseline and the shadowing conditions than did Experiment 1 [VOTs not extended; $t(14) = 1.95, p < .05$]. Among the difference scores, one score from each experiment was more than three standard deviations from the mean difference score. When these two outliers were removed from the analysis, Experiment 2 VOTs remained significantly longer than those in Experiment 1 [$t(12) = 2.87, p < .01$]. The same analysis was conducted, assuming unequal variances, and yielded the same significance level.

We asked whether the small increase in VOTs of the shadowed words in Experiment 2 reflected the influence of just a few tokens. That is, the participants showed just a 12-msec difference between baseline and shadowed words, whereas the lengthened model tokens of Experiment 2 had VOTs that were 68 msec longer than those of the original productions. Was the small size of the effect due to faithful imitations of just a few of the tokens? That was not the case. An items analysis performed on the baseline and shadowed VOTs of Experiment 2 was significant [$t(78) = 10.49, p < .0001$].

Another question concerned the basis for the VOT lengthening. Possibly, speakers lengthened the VOTs by slowing their word productions overall. This slowing would not be imitative of the models' words, which overall were longer than the original productions only by the difference in VOT. Slowing might be a response to the lesser goodness of the lengthened than of the original tokens. We measured the durations of the shadowing responses and the baseline responses. Shadowed words were longer than baseline words [455 msec, SD = 69, vs. 425 msec, SD = 59; $t(7) = 2.734, p < .05$]. We next asked whether the difference in duration itself differed significantly from the VOT difference between baseline and shadowed words. It did not approach significance. Baseline and shadowed words differed systematically in VOT and in total word duration. They did not differ systematically in word duration beyond the lengthening due to VOT.

Discussion

As was predicted, shadowed tokens were reported to be better imitations of target words (X) with extended VOTs than were baseline words. Unlike in Experiment 1, this occurred across the two AXB configurations (baseline token presented first vs. shadowed token presented first; see Table 1). In other words, the recency effect seen in Experiment 1 was absent this time. This was likely due to the fact that a particular property of the target word—namely, the VOT—was exaggerated in the target token and was, thus, a more salient property of the response item (Token A or B) for the participants to identify.

We found that VOTs were longer in the shadowing conditions of both experiments than in the baseline conditions. This may reflect the different conditions in which the tokens were collected. Baseline tokens were read, and shadowed tokens followed spoken words. Our uniform instructions to speak "quickly but clearly" may not have sufficed to eliminate differences in speaking style across the two conditions. Closer inspection of the data revealed, in any case, that on average, the model speakers' VOTs in Experiment 1 were longer than the baseline VOTs of the participants in that experiment. However, given that the model speakers' VOTs in Experiment 2 were extended, the difference between baseline and shadowed VOTs of Experiment 2 should be greater than the difference in Experiment 1 if imitation took place in both experiments. The latter outcome was, in fact, supported by a difference score comparison between the two experiments. Thus, the participants extended the VOTs of their shadowed responses when the target tokens were longer than their own baseline VOTs. This is particularly interesting because the participants were, in fact, modifying their articulatory dynamics to be more similar to the articulatory dynamics of the speaker models. Furthermore, this apparent gestural imitation was performed immediately, without instructions to do so.

GENERAL DISCUSSION

We have replicated Goldinger's (1998) finding that listeners imitate the speech that they hear, and we have taken a step toward addressing the kinds of speech properties that imitations preserve from the speech of a model speaker.

Following Goldinger (1998), we have found that, in essentially nonsocial settings in which there is no obvious motive to imitate, listeners/speakers nonetheless do so.

Goldinger was less interested in the fact that listeners/speakers imitate than in the conditions under which imitations increase or decrease in fidelity. His aim was to test an exemplar model of the mental lexicon having properties like those of the exemplar memory system proposed by Hintzman (e.g., 1986). In such a system, each utterance of the model speaker in Goldinger's research is stored as a distinct memory trace (a vector of features capturing the consonants and vowels of the utterance, the speaker's voice, and other information). When an utterance is perceived, traces in memory are activated to the extent that their features match those of the input utterance. Activated traces coalesce with the trace of the new input into an *echo* that Goldinger supposed influenced a speaker's shadowing response. Repeating utterances by a speaker in the familiarization phase of the experiment provided a way to shape the characteristics of the echo. Repetitions increased the number of identical traces in memory, which should make the echo for that utterance type increasingly resemble the repeated utterance in such properties as voice, speaking rate, and so forth, the more repetitions occurred. If the echo guides the shadowing response, imitative fidelity should increase with repetitions during familiarization, and in Goldinger's research, it did. Our manipulation of repetitions had no such effect in either Experiment 1 or 2, however. Accordingly, we did not replicate this evidence favoring the idea that lexical memory is an exemplar memory system.

In any case, our interest is not in the nature of lexical memory but, rather, in the observation that imitation occurs. When convergences in dialect, speaking rate, and other speech properties occur in social settings (e.g., Giles et al., 1991), they can be rationalized in various ways related to the idea that, in cooperative settings, speakers try to coordinate themselves with their interlocutors, or to get on their "wavelengths." This kind of account is supported by findings that divergences can occur in hostile encounters (Giles, 1973). However, the conditions of the experiments that we and that Goldinger conducted do not foster imitation on social grounds. Why else might imitation occur?

One account might be provided by the motor theory of speech perception (Liberman & Mattingly, 1985). In that theory, listeners perceive the linguistic vocal tract gestures of the speaker, and they do so in a way that involves their own speech motor system in the perceptual process. In this theory, because speech by others activates one's own speech motor system, speech can serve as a prime for an imitative response.

Another account might be provided by direct realist theory (e.g., Fowler, 1994), another gesture theory. It is that, in effect, listening to speech includes receiving instructions for the shadowing response. That is, if, as above, a listener hears a "high back positioning of the tongue with lips rounded" (produced at a particular rate with a particular emphasis and intonation contour), that can serve as instructions to produce a shadowing response, and if it does, the response is likely to be an imitation, because the information perceived will include more than just the information required to shadow. The direct realist account is consistent with the assumption that speech production and speech perception involve nonlinear dynamical systems that are self-organizing and reduce the potentially large number of degrees of freedom (to be controlled or detected) to a relatively small number of *order parameters* (e.g., Porter & Hogue, 1998). In both speech production and speech perception, it is assumed that the order parameters to be controlled or detected correspond to "those psychologically specified segments which, when combined in different ways, convey meaning" (Porter & Hogue, 1998, p. 97). The present research suggests that the order parameters in question may correspond to vocal tract gestures.

Although we think that findings of imitation are more compatible with a gestural than with an auditory theory, we acknowledge that accounts can be offered from that perspective as well. For example, perhaps imitation is such an important ability—in particular, in guiding children's entry into their cultural community—that humans are disposed to imitate quite generally. That is, as auditory theories propose, immediate perceptual objects are acoustic; individuals disposed to imitate attempt to match the acoustic signals that they hear (or the normalized acoustic signals).

Our research also begins to address the question of what is imitated. In Experiment 2, we lengthened the VOTs of our model speakers' word initial voiceless stops. We found evidence of imitations in listeners' judgments, as we had in Experiment 1, but we also found that our participants' shadowed productions had longer VOTs than did their baseline productions. Accordingly, one aspect of speech that is imitated is phonetic.

It will be of interest in future investigations to look at other properties of imitated speech. From an auditory theory's point of view, it is of interest to look at the fundamental frequency contour of shadowed speech. If imitations are based on acoustic perceptual objects, should we expect female speakers, say, to try to match the fundamental frequency of a male speaker, or do they, rather, attempt to match the location of the speaker's fundamental frequency in his pitch range with the corresponding location in their own pitch range (as if speech were normalized)? From the gesture theory's perspective, it will be useful to try to confirm specifically that gestural imitation is occurring. The research reported here is one study in a series that we have designed to uncover what listeners/speakers imitate, to uncover the basis (gestural or acoustic) for their imitations, and, perhaps, to expose the reasons why imitation is pervasive among humans.

REFERENCES

CHISTOVICH, L. A., KLASS, I. A., & KUZ'MIN, I. I. (1962). *The process of speech sound discrimination*. Boston: Emmanuel College Research Center.

COHEN, J., MACWHINNEY, B., FLATT, M., & PROVOST, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, & Computers*, **25**, 257-271.

CONDON, W. (1976). An analysis of behavioral organization. *Sign Language Studies*, **13**, 285-318.

FOWLER, C. A. (1994). Speech perception: Direct-realist theory. In R. E. Asher (Editor-in-Chief) & J. M. Y. Simpson (Coordinating Ed.), *The encyclopedia of language and linguistics* (Vol. 8, pp. 4199-4203). Oxford: Pergamon.

GILES, H. (1973). Accent mobility: A model and some data. *Anthropological Linguistics*, **15**, 87-105.

GILES, H., COUPLAND, J., & COUPLAND, N. (1991). Accommodation theory: Communication, context and consequences. In H. Giles, J. Coupland, & N. Coupland (Eds.), *Contexts of accommodation: Developments in applied sociolinguistics* (pp. 1-68). Cambridge: Cambridge University Press.

GOLDINGER, S. (1998). Echoes of echoes? Shadowing words and nonwords in an episodic lexicon. *Psychological Review*, **105**, 251-279.

HINTZMAN, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, **93**, 411-428.

KUČERA, H., & FRANCIS, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

KUHL, P. K., & MELTZOFF, A. N. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. *Journal of the Acoustical Society of America*, **100**, 2425-2438.

LAFRANCE, M. (1982). Posture matching and rapport. In M. Davis (Ed.), *Interaction rhythms: Periodicity in communicative behavior* (pp. 279-298). New York: Human Sciences Press.

LIBERMAN, A. M., COOPER, F. S., SHANKWEILER, D. P., & STUDDERT-KENNEDY, M. (1967). Perception of the speech code. *Psychological Review*, **74**, 431-461.

LIBERMAN, A. M., & MATTINGLY, I. G. (1985). The motor theory of speech perception revised. *Cognition*, **21**, 1-36.

LUCE, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

MCHUGO, G. J., LANZETTA, J. T., SULLIVAN, D. G., MASTERS, R. D., & ENGLIS, B. G. (1985). Emotional reactions to a political leader's expressive displays. *Journal of Personality & Social Psychology*, **49**, 1513-1529.

MELTZOFF, A. N., & MOORE, M. K. (1983). Newborn infants imitate adult facial gestures. *Child Development*, **54**, 702-709.

MELTZOFF, A. N., & MOORE, M. K. (1997). Explaining facial imitation: A theoretical model. *Early Development & Parenting*, **6**, 179-192.

MILLER, J. L., & VOLAITIS, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics*, **46**, 505-512.

PORTER, R. J., JR., & CASTELLANOS, F. X. (1980). Speech production measures of speech perception: Rapid shadowing of VCV syllables. *Journal of the Acoustical Society of America*, **67**, 1349-1356.

PORTER, R. J. [JR.], & HOGUE, D. M. (1998). Nonlinear dynamical systems in speech perception and production. *Nonlinear Dynamics, Psychology, & Life Sciences*, **2**, 95-131.

PORTER, R. J. [JR.], & LUBKER, J. F. (1980). Rapid reproduction of vowel–vowel sequences: Evidence for a fast and direct acoustic–motoric linkage in speech. *Journal of Speech & Hearing Research*, **23**, 593-602.

SHOCKLEY, K., SANTANA, M. V., & FOWLER, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception & Performance*, **29**, 326-332.

NOTE

1. We used these instructions, following Goldinger (1998), in instructing both our model speakers and our participant speakers. The purpose was to get productions by the model speakers and the participant speakers that were similar in overall rate and speaking style.

APPENDIX A
English Words and Frequencies

/k/	Frequency (per Million)	/p/	Frequency (per Million)	/t/	Frequency (per Million)
cabbage	4	package	20	tailor	2
cable	7	painter	21	tamper	1
camel	1	panic	22	target	45
campus	33	panther	1	taxi	16
canyon	12	pardon	8	teaspoon	4
capture	17	parrot	1	temper	12
carpet	13	partner	32	temple	38
cartridge	6	passion	28	tender	11
castle	7	patience	22	tennis	15
caution	19	payment	53	terrace	9
cocoa	2	pedal	4	ticket	16
combat	27	pencil	34	tidy	1
comet	2	penny	25	tiger	7
compass	13	perfect	58	timber	19
concert	39	pester	1	timing	11
contact	63	pickle	1	tissue	41
contest	26	pigeon	3	token	10
copper	13	pillow	8	tonic	1
costume	10	pizza	3	topic	9
cottage	19	poison	10	towel	6
courage	32	poker	6	tuba	1
culture	58	poodle	2	tulip	4
curtain	13	poster	4	tumble	3
cushion	8	posture	13	tunnel	10
custom	14	punish	3	turkey	9
kennel	3	puppy	2	turtle	8
kitten	5	puzzle	10		
M	17.3		14.6		11.9

Note—Frequencies of words are based on Kučera and Francis (1967).

APPENDIX B

Editing Target Tokens in Experiment 1

The model speakers' tokens from Experiment 1 were modified, using SoundScope (GW Instruments, Somerville, MA), by extending the VOT of the initial consonant for each word. The original VOT for the initial consonant was measured, and medial portions of the aspiration were selected, copied, and pasted back into the aspiration section of the waveform. Duration and placement of selections varied across and within tokens. For example, the VOT for /p/ in the female unedited token *painter* was 48 msec. The extended VOT of 96 msec was obtained by making an initial selection of 9.06 msec, and subsequent selections of 12, 8.75, 9.06, and 8.75 msec. These durations varied because we had to be careful to select stretches of VOT that, pasted back into the utterance, did not lead to perception of clicks or other evidence of the splicing operation.

The average overall extension of VOTs for edited tokens was 2.03 times that of the original VOT values of the unedited tokens ($SD = 0.228$). Editing the female model's tokens created extended word-initial VOTs that were 2.02, 2.01, and 2.03 times longer than the original VOTs for /p/, /t/, and /k/, respectively [$F(2,76) = 1.68$, $p = .19$]. Editing the male model's tokens created extended word-initial VOTs that were 2.12, 2.00, and 2.02 times longer than the original VOTs for /p/, /t/, and /k/, respectively ($F < 1$).

The edited tokens were saved as files for use in PsyScope (Cohen, MacWhinney, Flatt, & Provost, 1993). The mean measured VOTs before extension for consonants /k/, /p/, and /t/ were 64.4 ($SD = 10.7$), 61.9 ($SD = 15.5$), and 62.8 ($SD = 10.1$) for the female model and 78.1 ($SD = 14.9$), 57.7 ($SD = 22.3$), and 77.2 ($SD = 13.9$) for the male model, respectively.

Measuring VOT for Responses in Experiment 2

The participants' responses were recorded onto a cassette tape and later were input to SoundScope (see above). We measured VOTs of the word-initial consonants using both the waveforms and the spectrographic displays of the participants' responses. Duration of VOT was measured from the beginning of the release burst to the beginning of glottal pulsing.

The beginning of the VOT was designated as the leftmost edge of the release burst. Any noise prior to the highest energy point in the burst was assessed relative to the major portion of apparent speech energy in the wave. If the noise seemed to be part of the burst in terms of distance and quality, as with some productions of /k/, which often had several small release bursts that were quite close to one another and to the major burst, the noise was counted as part of the VOT. If the noise seemed to be independent of the burst on the basis of distance and quality, such as a lip smack before a /t/ or a slight cough before a /p/, it was excluded from the measurement.

The end of the VOT was the point in the wave just before voicing was evident. Investigators discerned evidence of voicing aurally and visually. Visual information consisted of signs of the onset of glottal pulsing in the waveform under high magnification, as well as the appearance of a voicing bar in the spectrogram.

(Manuscript received February 20, 2002;
revision accepted for publication August 18, 2003.)