PRODUCING RELATIVELY UNFAMILIAR SPEECH GESTURES: A SYNTHESIS OF PERCEPTUAL TARGETS AND PRODUCTION RULES

G. J. Borden,* K. S. Harris,** Hollis Fitch,*** and H. Yoshioka****

Abstract. Attempts of speakers to imitate familiar and foreign syllables under adverse feedback conditions were analyzed by perceptual judgments, electromyographic recordings, and spectrographic measures. Although foreign syllables were more poorly imitated than familiar syllables, decrements in feedback interfered more with familiar than with novel utterances. Decrements in acoustic, tactile, and proprioceptive information were worse in combination than singly. Speakers did not improve unfamiliar fricative production under any condition on 13 learning trials.

Research during the last decade has demonstrated that intelligibility of the speech of skilled speakers remains high despite masking of the speakers' auditory feedback or decreasing their tactile feedback. There is some segmental distortion when the tongue is anesthetized (Ringel & Steer, 1963; Borden, Harris, & Oliver, 1973), and some prosodic distortion when speech is attempted under simultaneous but modified auditory feedback (Lane & Tranel, 1971; Siegel & Fick, 1974), but the overall effect upon speech production seems to be surprisingly small (see Borden, 1979, for a review).

These findings argue for the importance of a feedforward system for production of well-known motor patterns for speech, with auditory and tactile information used for fine tuning or correction of errors. The adult speaker seems to know the possibilities of his or her own vocal tract. Simple constraints on movement imposed by talking with a pipe or pencil clenched between the teeth or with an experimental bite block do not alter the vocal tract dimensions, and interference with speech is minimal (Lindblom & Sundberg, 1971). These results are consonant with those of animal experiments, in which direct and complete elimination of sensory information is accomplished by surgical means. It has been shown that monkeys trained to perform specific movements can continue to do so, despite deprivation of feedback from limbs or chewing muscles (Taub & Berman, 1968; Goodwin & Luschei, 1974; Polit & Bizzi, 1978), although there are indications that new movements are impaired (Polit & Bizzi, 1978).

*Also Temple University, Philadelphia, PA.
**Also City University of New York, New York, NY.
***Also Institute for Defense Analyses, Princeton, NJ.
****Also University of Tokyo, Tokyo, Japan.

Acknowledgment. The writing of this article was partially supported by NIH grant NS13870 to Haskins Laboratories. We are grateful to Arthur Abramson for taping the model stimuli and for judging the subjects' utterances under normal conditions, and to Anders LÖfqvist for serving as the expert listener.

Of course, we cannot know with certainty the role of self-monitoring of speech, because it is impossible to eliminate simultaneously all channels of information available to a speaker. We do know that sensory information is important in learning speech for the first time or in successfully learning the speech patterns of a new language. The labored speech of the deaf (Osberger & Levitt, 1979; Harris & McGarr, 1980) and the rare case of a speaker with an oro-sensory loss (Chase, 1967) testify to the importance of self-monitoring while learning speech. We know, too, that normal adult speakers often need time to adjust to prosthetic devices that alter the dimensions of their vocal tracts (Hamlet & Stone, 1976), and feedback of auditory, tactile, and proprioceptive information is presumed to control the compensatory patterns that evolve.

There have been no studies to our knowledge, however, that investigate self-monitoring of speech by comparing the effects of diminished sensory information on the performance of speech gestures new to the speaker with those familiar to the speaker, with the exception of one report to the effect that children who are better than other children at identifying forms placed in the mouth (oral stereognosis) are also better at learning non-native speech sounds (Locke, 1968).

In the present investigation, we were interested in exploring whether perceptually accurate speech sounds would be produced under conditions of adverse speech control when the speech gestures were not those learned as part of the language of the speaker. How well might the speaker control production of non-English syllables? Might vowels and consonants depend differently upon sensory information? How well might the speaker control English and non-English utterances when auditory feedback is diminished? when tactile information is decreased? when vocal tract configuration is altered? The question that motivated these experiments was not what happens to speakers with loss of feedback—the speaking conditions reported in this paper represent diminished or altered feedback, not its absence—rather, the question is how do familiar versus relatively novel speech gestures hold up under various conditions and combinations of conditions that alter or diminish information that is normally fed back to the speaker as he is talking?

Two approaches can be taken to judge adequacy of performance. One approach is to measure some aspect of production directly in various conditions—here we have measured articulator activity using EMG techniques, and some aspects of acoustic output, using conventional spectrographic analysis. Another approach is to examine perceptual adequacy by using listener judgments of performance. The second approach has the disadvantage of being subjective, but does measure communicative adequacy. While the first approach is objective, any particular set of measurements is not exhaustive.

One can rationalize three hypotheses about the experimental outcome: The first is that relatively novel utterances will suffer more than familiar utterances under conditions of altered or diminished information, because speakers might need more information for the less familiar utterances. The second hypothesis is that familiar utterances would suffer more than novel under deprived feedback conditions, because speakers might hold internalized finely developed auditory-ororo-sensory criteria for the well-learned utterances and might use feedback to sharpen the match between their utterances and these
criteria. For the less familiar utterances, however, speakers may hold only broad criteria for how the speech should sound and feel, and therefore make less use of information from the periphery. The third hypothesis is that familiarity would make little difference, because speakers might not succeed in producing unfamiliar motor sequences even when all feedback information is available; they might convert less familiar utterances into familiar ones and utter a variant of a similar sound from their own language system.

PROCEDURE: PRODUCTION TASK AND PERCEPTUAL ANALYSIS

The general design of the investigation was to have subjects imitate a recording of a phonetician saying syllables that were within the phonetic inventory of English and syllables that were phonetically foreign to English. The speakers imitated the speech sounds under normal speaking conditions and under altered speaking conditions: auditory masking, lingual anesthesia, and alterations of the shape of the palatal vault. The speech was recorded acoustically and the muscle activity of the tongue was analyzed by electromyographic measures. Tapes for each speaker made by pairing utterances spoken by the phonetician with utterances spoken by a subject under various speaking conditions were used for perceptual tests to assess the judged differences between speaking conditions.

Subjects

Three normal adult males served as the primary subjects for the experiment. They were speakers of American English, and, although they had studied languages other than English in school, each subject was essentially monolingual with little practical experience in speaking any other languages. Two of the subjects were 21 years old (DB and GF) and the third was 33 (TB). None was informed of the purpose of the experiment. Since the long-lasting effects of the anesthesia condition precluded the perfect balancing of orders, four additional subjects were recorded each with a different order of conditions. These speakers were run without nerve-block anesthesia of the tongue and without electromyographic insertions to see what order effects there might be, and to enlarge the subject pool. The non-nerve-block speakers were students at Temple University and were also naive about the purpose of the investigation. As other subjects were used for the perceptual part of the analysis, we shall avoid confusion by referring to the imitators as speakers and to the subjects of the perceptual tests as listeners.

Speech Task

For this investigation we chose a small set of speech sounds, some that would be familiar to monolingual speakers of American English and some that would be relatively novel. The criteria were that the sounds must exist in some language and they must be acoustically distinct. We chose two familiar vowels [i] and [e\textsuperscript{2}] (as in 'see' and 'say) and two familiar consonants, one voiceless [f] and one voiced [z] (as in 'shoe' and 'zoo'). For the less familiar sounds the vowels [y] and [\textsuperscript{3}\delta] (as in the French words 'tu' and 'deux') were chosen because they are rounded front vowels not phonologically
present in English. The novel consonants chosen were the voiceless and voiced palato-velar fricatives [x] and [ɣ] (as in the Spanish words 'rojo' and 'rogar'). The vowels were initiated with [p] and the fricatives were followed by [i] yielding eight syllables. A phonetician proficient in the production of all eight of these speech sounds recorded them in syllable form after the word 'say.' The list was read three times and a satisfactory token of each type was chosen to be digitized on the Haskins PCM system. A tape recording was constructed containing a list of 24 utterances (each utterance repeated three times and randomized) followed by eight lists in which each syllable type was repeated 10 times. The last eight repetition lists were used to investigate learning.

Experimental Conditions

The three primary speakers from whom electromyographic data were collected were recorded under conditions of auditory masking, lingual anesthesia, false palate, and combinations of these conditions as well as the normal speaking condition used as a control. The four speakers recorded without EMG insertions were recorded in the same conditions as the primary subjects with the exception of the condition of lingual anesthesia.

The condition of diminished auditory feedback was achieved by recording the speech of the phonetician on one channel and white noise on the second channel of a tape recording. The speech was delivered binaurally at 70 dB SPL and the white noise, also binaurally, at 90 dB SPL during the subjects' responses. To control vocal intensity, subjects were instructed to monitor the VU meter on the tape recorder that was recording their responses: they were not to let their vocal intensity rise above the midpoint of the range, representing about 55 or 60 dB. Although the low frequencies of the voice were undoubtedly transmitted by bone conduction, the higher frequency contribution of the vocal tract resonances to the various speech sounds was essentially masked for the speaker.

Linguual anesthesia was produced by blocking the sensory fibers of the lingual nerve on both sides of the jaw. The lingual nerve, a branch of the Trigeminal nerve, was blocked by a dentist who bilaterally injected 1.8 cc of 3 percent Carbocaine containing a vasoconstrictor. The criterion for lingual anesthesia was that the superior surface of the anterior two-thirds of the tongue must be insensitive to a dental probe.

The conditions of masking noise and nerve block resulted in diminished auditory and tactile feedback, respectively. Proprioceptive feedback, in this case information on tongue position and movement, is impossible to interrupt short of surgical techniques. To impoverish the usefulness of the proprioceptive information, however, the shape of the vocal tract was altered by placing a dental impression material, Alginate, on the superior alveolar ridge behind the central and lateral incisors. The material extended posteriorly along the hard palate for several centimeters. Whatever proprioceptive information the speaker may have received from tongue position and movement within the vocal tract, the fact that vocal tract volume was changed, thus altering the presumed coordinates of the space, would alter the customary reference points for proprioceptive information. After the impression material was removed
from the mouth of each subject, it was cut along a central line extending from between the central incisors, and the width of the portion corresponding to the apex of the alveolar ridge was measured. The addition of material resulted in a buildup of the ridge by 6 mm for each subject.

Conditions were applied singly and in combination with orders varied. For the primary speakers, speaking conditions were given in the following order. M stands for auditory masking, NB for the nerve block resulting in lingual anesthesia, and A for Alginate, the dental impression material used to alter the architecture of the palate. For DB, the order was NB, M, and A; NB and A; and finally, NB alone. For TB, the order was NB; NB and M; NB, M, and A, and NB and A. For GF, the order was M; M and A; A; and M, A, and NB. The control condition was recorded on another day to ensure that there were no effects of the anesthesia. For the non-nerve-block subjects four orders were possible reserving the control condition for last: 1) A; A and M; M, 2) M; A; A and M; A, and 4) A and M; A; M. The order A and M; M; A was not possible as the impression material could not be removed and reinserted.

Electromyographic Recording

Hooked wire electrodes of .002 inch platinum alloy were inserted into the superior orbicularis oris muscle (00), the superior longitudinal muscle (SL), the inferior longitudinal muscle (IL), and the genioglossus muscle (GG) of the three primary subjects. The orbicularis oris muscle was sampled to allow for observations of muscle activity in the lips for the rounded less familiar vowels. The genioglossus muscle was sampled to assess production of the high front vowels, and the intrinsic muscles of the tongue were sampled in an effort to observe differential tongue activity for production of the the fricatives. The EMG recordings consisted of the eight speech task utterance types, 13 tokens of each, under all speaking conditions. Only the EMG signals recorded during the three tokens of each type in the randomized list of 24 items have been analyzed. The signals were rectified, smoothed with a 35 msec time constant, and digitized. Procedures for insertion, recording, and analysis are described in detail elsewhere (Hirose, 1971; Kewley-Port, 1973).

Acoustic Recording

Sound spectrograms were made of all utterances spoken by the primary speakers. Second formant frequencies were measured for [i], [eI], [y] and [ø]. Normally, the rounded front vowels /y/ and /ø/ are realized acoustically with higher F1, and lower F2 than the unrounded front vowels /i/ and /eI/ (Pols, Tromp, & Plomp, 1973). The tongue is thought to be higher for the unrounded members of the respective pairs /i-y/ and /eI-ø/ (Raphael, Bell-Berti, Collier, & Baer, 1979).

The fricative consonants were measured in the center of the third formant noise. Normally, the prominent resonances for [s] are lower in frequency (approximately 2500 Hz) than those for [z] (approximately 4000 Hz). Figure 1 contrasts the [s] and [z] resonances in the model "Say [ iJ" and "Say [zi]." Figure 2 shows the spectrographic representation of the utterance "Say [Xi]" and "Say [yi]" as spoken by the phonetician used in this study. F2 and
Model Phonetician

Figure 1
Model Phonetician

Figure 2
F₃ are close together for [X] and [γ] with F₃ ranging between 2000 Hz and 2500 Hz in an average male vocal tract. Conspicuous is the antiresonance below the second formant. In cases where fricative energy was low in the F₃ region, the F₃ frequency at the onset of the following vowel was measured.

Perceptual Testing

A listening tape was constructed from the model utterances of the phonetician and the imitations of each subject (one tape for each speaker) by digitizing all of the speech samples and editing them on the Haskins PCM system so that for each syllable type, each speaking condition was contrasted with each of the other speaking conditions in both orders. Each trial presented the model utterance, for example "Say zi" as said by the phonetician, followed by the speaker's imitation under one condition, then the phonetician again, followed by the speaker's imitation under another condition. The phonetician's utterance and the imitations were separated by 500 msec, and the pairs for each trial were separated by 1500 msec. A 3 second pause between trials allowed time for listeners to check on answer sheets the imitation they preferred. With five conditions (yielding 10 condition contrasts and with orders of pairs reversed, 20 condition contrasts) of 24 utterances (8 types, 3 tokens each), each listening test consisted of 24 lists of 20 trials each for a test of 480 items. Trials were randomized throughout each test, and each condition was paired with every other condition with orders reversed. Each test was divided into two tapes. Listeners were 27 students from the University of Connecticut, 9 to judge the two test tapes for one of the three speakers. Each tape took approximately one hour. Listeners were asked to judge pronunciation and to disregard any change in loudness or pitch. They were to indicate which of the two imitations in each contrast "more successfully matched the speech sounds" of the phonetician. For the tape constructed from the responses of the first speaker, judgments of three expert listeners were collected to compare with the judgments of the relatively naive student listeners, to assess the effects of listener perceptual sophistication.

Listening tapes were also constructed from the responses of four speakers who did not receive a nerve block. Again, students from the University of Connecticut served as listeners. The listeners were instructed to mark the imitation judged worse than the other with a check and, if much worse, with an X. This change in procedure was an attempt to obtain an idea of the relative magnitude of decrement in perceived pronunciation resulting from the experimental conditions.

RESULTS

Analysis of the data can be divided into the electromyographic analysis, spectrographic analysis, and the analysis of listener judgments. We shall briefly mention the EMG and spectrographic results first, and devote more space to the perceptual results.
Electromyographic Analysis

The first three samples of each syllable type, spoken under each condition, have been analyzed for the three primary speakers. Peak amplitude measures for each electrode placement were graphed. Timing measures were also made.

The muscle that is the primary contributor to lip closure and lip rounding, the orbicularis oris muscle (OO), was active for all three speakers during the rounded vowels [y] and [ø], while it was inactive for [i] and [eI]. Figure 3 shows the contrast between the two types. There is a compact peak of activity for the [p] in [pi] starting before the vertical line at zero. The line indicates the termination of the vowel in 'Say' for the utterance 'Say [pi].' The [p] for [py] is also preceded by a compact burst of muscle activity, but OO remains fairly active (324V at around 400 msec) throughout the vowel. All these speakers showed evidence of OO activity for the unfamiliar vowels [py] and [pø].

Successful recordings were made from GG for two speakers, and were examined for productions of [i]. Activity was remarkably stable for TB, especially the timing of the activity (Figure 4). Peak amplitude was lower with the addition of Alginate. For speaker GF, GG activity for [i] tended to be more diffuse and drawn out as the speaking conditions got more complicated.

The patterns of activity for SL and IL differed from subject to subject. In general, when either muscle was active for a given fricative, the activity often became erratic with the application of Alginate to the palate, with an increase in activity recorded from IL for two of the speakers. IL normally depresses the tip of the tongue. One speaker (TB) showed little change in SL for [z] in the Alginate condition, but showed a decrease of IL activity (Figure 5). Since TB produces [s] and [z] with the tip of the tongue curled down behind the lower incisors bunching the dorsum of the tongue for the constriction (Borden & Gay, 1979), we assume that the pattern represents a decrease in bunching.

Only one speaker (DB) used SL for fricatives other than [z], limiting comparisons between novel and familiar consonants to that speaker. Comparing the electromyograms of [z], the least variable fricative for DB, with the most variable, [X], activity recorded from SL in the worst speaking condition (nerve-block, alginate, and auditory masking) remained essentially the same for the tokens of [z] but varied considerably for tokens of [X] (Figure 6). The first utterance was transcribed as [z] in all instances, but the second utterance was transcribed as [Xr], a velar fricative with a retroflexed tongue, in the first imitation and as [ʃ], a voiced pharyngeal fricative, for the second imitation. SL was active for the better imitation but completely inactive for the pharyngeal fricative.

In general, electromyographic recordings confirmed the observations described below. First, speakers imitated the unfamiliar rounded front vowels adequately in all conditions, using lip rounding to do so. Second, adverse speaking conditions tended to result in reduced tongue activity or erratic patterns.
ORBICULARIS ORIS MUSCLE

Figure 3
Figure 4
NERVE BLOCK

NERVE BLOCK AND ALGINATE

\[ \mu V \]

TB

SLm

ILm

"say zi"

"say zi"

100
SUPERIOR LONGITUDINAL MUSCLE FOR [Z] AND [X]

Two Trials of Each Syllable Type

[Z] Imitation

[X] Imitation

Figure 6
Table 1

$F_2$ for Vowels

<table>
<thead>
<tr>
<th></th>
<th>$i$</th>
<th>$y$</th>
<th>$e^\pi$</th>
<th>$\phi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>S:DB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>2233</td>
<td>47</td>
<td>1683</td>
<td>62</td>
</tr>
<tr>
<td>NB</td>
<td>2133</td>
<td>24</td>
<td>1817</td>
<td>131</td>
</tr>
<tr>
<td>NB + M</td>
<td>1900</td>
<td>41</td>
<td>1600</td>
<td>141</td>
</tr>
<tr>
<td>NB, M, &amp; A</td>
<td>1867</td>
<td>47</td>
<td>1733</td>
<td>262</td>
</tr>
<tr>
<td>S:TB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1783</td>
<td>24</td>
<td>1300</td>
<td>41</td>
</tr>
<tr>
<td>NB</td>
<td>1767</td>
<td>23</td>
<td>1416</td>
<td>62</td>
</tr>
<tr>
<td>NB + M</td>
<td>1700</td>
<td>0</td>
<td>1400</td>
<td>62</td>
</tr>
<tr>
<td>NB + A</td>
<td>1700</td>
<td>41</td>
<td>1325</td>
<td>25</td>
</tr>
<tr>
<td>NB, M, &amp; A</td>
<td>1600</td>
<td>71</td>
<td>1233</td>
<td>47</td>
</tr>
<tr>
<td>S:GF</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1967</td>
<td>47</td>
<td>1683</td>
<td>24</td>
</tr>
<tr>
<td>M</td>
<td>1733</td>
<td>24</td>
<td>1467</td>
<td>112</td>
</tr>
<tr>
<td>A</td>
<td>1858</td>
<td>24</td>
<td>1683</td>
<td>43</td>
</tr>
<tr>
<td>M &amp; A</td>
<td>1667</td>
<td>23</td>
<td>1383</td>
<td>24</td>
</tr>
<tr>
<td>NB, M, &amp; A</td>
<td>1717</td>
<td>47</td>
<td>1558</td>
<td>83</td>
</tr>
</tbody>
</table>
Spectographic Analysis

Measurements of the second formants of the vowels for the three primary speakers in this study are presented in Table 1. The means are based on the three tokens of each syllable under each speaking condition that was used in the perceptual test. The acoustic difference between pairs found in previous studies (Pols et al., 1973; Raphael et al., 1979) holds under all speaking conditions for the [i]-[y] contrast: F2 for [i] is higher than for [y]. The difference is maintained for [eɪ] and [ʊ] under normal speaking conditions, but does not hold under all adverse conditions. A prominent effect upon the formant frequencies of the vowels is seen in the condition of auditory masking. Generally, when subjects are prevented from hearing the higher resonances of their voices during front vowel production, the resonances drop in frequency somewhat. Also, there is a tendency for variability to be greater for the F2 of the less familiar vowels [y] and [ʊ] than for the familiar vowels [i] and [eɪ].

Table 2 details the means and standard deviations of the F3 resonances for the fricatives as imitated by the three speakers under various speaking conditions. For speaker GF the condition involving alginate on the alveolar ridge resulted in lower vocal tract resonances in the F3 region than for other conditions, but the other two speakers showed little effect. Variability was apt to be higher on unfamiliar syllables and during combined deprivation conditions but not consistently so.

Spectrograms of the imitations of DB’s [z] and [X] utterances corresponding to the EMG plots shown in Figure 5 are shown in Figures 7 and 8. Figure 7 (a and b) represents a wide band and a narrow band display of two imitations of "Say zi" as produced under the combined condition of alginate, nerve block, and auditory masking. Figure 8 shows two imitations of [X]. The first attempt (Figure 8a) consists of fricative noise, but the formants decline in frequency. It was transcribed as [hr] and as [XR] due to its liquid quality. The second attempt (Figure 8b) consists of fricative noise, but voicing continues and it was transcribed as a pharyngeal fricative [ɭ] or as a voiced aspirate. Again, note the difference in superior longitudinal muscle activity in Figure 5.

The spectrograms of the 10 repetitions of the fricative syllables [X] and [ɭ] for each speaker under each condition were also measured. Figure 9 represents the plots for each speaker of F3 frequencies across 13 trials (10 repetitions and the 3 tokens in the initial list). There is no systematic change that would indicate the presence of learning.

Perceptual Analysis

The purpose of obtaining listener judgments of the speaker imitations was to investigate the perceptual effects of the various speaking conditions on speakers' ability to imitate the familiar and relatively unfamiliar utterances. It was obvious that the familiar syllables were closer to the model under all conditions than were the unfamiliar syllables.
Table 2

$F_3$ for Fricatives

Based on Three Trials Each

<table>
<thead>
<tr>
<th>S:DB</th>
<th>$\tilde{\chi}$</th>
<th>$\tilde{\chi}$</th>
<th>$\chi$</th>
<th>$\chi$</th>
<th>$\gamma$</th>
<th>$\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>2567</td>
<td>23</td>
<td>2567</td>
<td>143</td>
<td>2500</td>
<td>41</td>
</tr>
<tr>
<td>NB</td>
<td>2583</td>
<td>131</td>
<td>2483</td>
<td>165</td>
<td>2683</td>
<td>24</td>
</tr>
<tr>
<td>NB + M</td>
<td>2450</td>
<td>122</td>
<td>2183</td>
<td>24</td>
<td>2533</td>
<td>24</td>
</tr>
<tr>
<td>NB, M, &amp; A</td>
<td>2667</td>
<td>47</td>
<td>2200</td>
<td>349</td>
<td>2683</td>
<td>24</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>S:TB</th>
<th>$\tilde{\chi}$</th>
<th>$\tilde{\chi}$</th>
<th>$\chi$</th>
<th>$\chi$</th>
<th>$\gamma$</th>
<th>$\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>2416</td>
<td>23</td>
<td>2567</td>
<td>47</td>
<td>2267</td>
<td>23</td>
</tr>
<tr>
<td>NB</td>
<td>2483</td>
<td>24</td>
<td>2400</td>
<td>41</td>
<td>2267</td>
<td>23</td>
</tr>
<tr>
<td>NB + M</td>
<td>2450</td>
<td>147</td>
<td>2367</td>
<td>62</td>
<td>2333</td>
<td>47</td>
</tr>
<tr>
<td>NB + A</td>
<td>2467</td>
<td>125</td>
<td>2667</td>
<td>47</td>
<td>2367</td>
<td>47</td>
</tr>
<tr>
<td>NB, M, &amp; A</td>
<td>2500</td>
<td>0</td>
<td>2267</td>
<td>47</td>
<td>2500</td>
<td>71</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>S:GF</th>
<th>$\tilde{\chi}$</th>
<th>$\tilde{\chi}$</th>
<th>$\chi$</th>
<th>$\chi$</th>
<th>$\gamma$</th>
<th>$\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>2425</td>
<td>54</td>
<td>2500</td>
<td>204</td>
<td>2333</td>
<td>24</td>
</tr>
<tr>
<td>M</td>
<td>2375</td>
<td>0</td>
<td>2442</td>
<td>51</td>
<td>2208</td>
<td>59</td>
</tr>
<tr>
<td>A</td>
<td>2200</td>
<td>35</td>
<td>2133</td>
<td>103</td>
<td>2125</td>
<td>0</td>
</tr>
<tr>
<td>M &amp; A</td>
<td>2075</td>
<td>54</td>
<td>1867</td>
<td>92</td>
<td>2050</td>
<td>54</td>
</tr>
<tr>
<td>NB, M, &amp; A</td>
<td>2333</td>
<td>118</td>
<td>2142</td>
<td>23</td>
<td>2257</td>
<td>12</td>
</tr>
</tbody>
</table>

100
Figure 7a
Figure 8a
SPECTROGRAPHIC VARIATION ACROSS TRIALS

Figure 9
Effects of single versus combined sorts of decrement in information. Figure 10 collapses listener judgments of seven speakers—the three speakers with nerve block and the four non nerve-block speakers. This figure represents averaged listener comparisons of altered speaking conditions with the normal condition. Listeners preferred the normal speaking condition to between 60-65 percent of the utterances spoken during any single alteration. The normal condition was preferred to between approximately 80-95 percent of the combined conditions. Thus, decrements in information available to the speaker, although of different sorts, impair speech more in combination than in any single condition.

Effects of speaking condition on familiar versus unfamiliar syllables. To look more closely at how the speaking conditions affect judgments of familiar versus less familiar utterances and judgments of vowel versus fricative syllables, we ran an analysis of variance on all possible paired contrast conditions from the perceptual data for each of the seven speakers. Two within-subject variables were explored: familiarity (familiar versus novel) and syllable type (vowel versus consonant). It can be seen in Table 3 that there is an effect of familiarity for some of the subjects, while there is an effect of syllable type for only one speaker. In some cases there is an interaction of familiarity with syllable type. For three speakers there was no perceptual effect of familiarity, syllable type, or their interaction that reached significance.

In all cases in which there was a significant effect of familiarity, listeners reported stronger perceptual differences between familiar syllables spoken under contrasting conditions than between less familiar syllables spoken under the same contrasting conditions.

In the cases in which there was a significant interaction between familiarity and syllable type, the interaction was speaker-specific. For TB, the novel consonants were perceived to be least changed. This confirms the general perceptual impression that [X] and [y] were rather consistently produced as [f] and [3] despite speaking condition. For the non-nerve-block speaker representing the second order of conditions, S2, the familiar consonants were perceptually more affected than the other syllable types, while for the speaker representing order S4, it was the novel vowels [y-Ø] that were perceived to be least changed by conditions.

In summary, differences in imitations of familiar and less familiar vowels and fricatives are more marked according to listener judgments in familiar syllables than unfamiliar ones, and the interaction between familiarity and syllable type depends upon the speaker.

Expert versus student listeners. A possible explanation of the "familiarity" effect is that it is due to the differences in listener familiarity with the sounds, rather than to differences in the productions themselves. In order to examine this possibility, we compared the judgments of an expert listener, naive to the purposes of the study, with the judgments of the student listeners. The expert listener was in general agreement with the naive student listeners in his judgments of relative deterioration of imitations under various speaking conditions. He too found larger differences among the familiar utterances, even though to him all the utterances were more
Normal > 

\[
\begin{align*}
&\text{Masking (60\%)} \\
&\text{Nerve Block (63\%)} \\
&\text{Alginate (65\%)} \\
&\text{Nerve Block, Masking & Alginate (81\%)} \\
&\text{Masking & Alginate (82\%)} \\
&\text{Nerve Block & Masking (83\%)} \\
&\text{Nerve Block & Alginate (86\%)}
\end{align*}
\]
Table 3

Analysis of Variance of Perceptual Data Collected on 7 Speakers
Effects of Familiarity, Syllable Type, and Their Interaction

<table>
<thead>
<tr>
<th></th>
<th>FAM.</th>
<th>SYLL.</th>
<th>INT.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB</td>
<td>F=21.8</td>
<td>F=23.2</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>p&lt;.001</td>
<td>p&lt;.001</td>
<td></td>
</tr>
<tr>
<td>Primary Speakers (df 1,9)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TB</td>
<td>F=10.6</td>
<td></td>
<td>F=5.1</td>
</tr>
<tr>
<td></td>
<td>p&lt;.005</td>
<td></td>
<td>p&lt;.05</td>
</tr>
<tr>
<td>GF</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S1</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>S2</td>
<td>NS</td>
<td>NS</td>
<td>F=7.967</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>p&lt;.05</td>
</tr>
<tr>
<td>Speakers Without Nerve Block (df 1,5)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>S4</td>
<td>F=31.506</td>
<td>NS</td>
<td>F=11.7</td>
</tr>
<tr>
<td></td>
<td>p&lt;.001</td>
<td></td>
<td>p&lt;.01</td>
</tr>
</tbody>
</table>
familiar than they were to the student listeners. Figure 11 contrasts typical response plots of the expert listener and the student listeners for a familiar and an unfamiliar syllable. For the [ʃ], listener scores reflected a decrease in listener preference as the speaking conditions got more complex, especially upon the addition of auditory masking and alginate, while for [γ], scores varied less from one condition to another.

The phonetician who served as the model speaker listened to the lists of 24 utterances spoken under normal conditions, (3 x 8 tokens) by the three primary subjects and the four additional subjects, and judged each imitation to be 1) Americanized or American, 2) Almost Americanized, 3) Neither American nor foreign, 4) Almost foreign, or 5) Authentic Foreign Accent. All speakers were judged to produce ordinary American productions of all familiar syllable types.

Table 4 summarizes for all seven speakers the percentages of familiar and less familiar utterances judged by the phonetician to be correct. The familiar utterances are counted as correct if they are judged to fall within the American English system. The less familiar utterances are counted as correct if they are judged to be within the foreign sound system.

Table 4
Judgments Made by Phonetician of Utterances Spoken Under Normal Speaking Conditions by 7 Speakers

<table>
<thead>
<tr>
<th>Familiar Utterances Judged Within American English System (Correct)</th>
<th>Less Familiar Utterances Judged Within Foreign Sound System (Correct)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. [ʃ] 100%</td>
<td>1. [φ] 33%*</td>
</tr>
<tr>
<td>2. [eɪ] 90%</td>
<td>2. [χ] 29%+</td>
</tr>
<tr>
<td>3. [i] 81%</td>
<td>3. [γ] 5%+</td>
</tr>
<tr>
<td>4. [z] 75%</td>
<td>4. [y] 0%*</td>
</tr>
</tbody>
</table>

N = 21
7 speakers x 3 tokens

*Never Americanized; 57% Judged Almost Foreign
+Americanized; 14% [χ] and 29% [γ]
LISTENER JUDGMENTS

S: DB

- EXPERT
- STUDENTS (MEAN)

[fi]

NUMBER OF UTTERANCES PREFERRED

[yi]

CONDITIONS

Figure 11
Even under normal speaking conditions, [ʃ] was judged to be acceptable despite acoustic variation, while [z] was more vulnerable to perceptual inconstancy. American students were apt to Americanize the novel fricatives, but never Americanized the rounded front vowels, which were judged more than half the time to be almost foreign or correct in production.

Effects of conditions on syllable types. The nerve block condition results in a rather small effect across syllables (Figure 12). A 50% preference would be expected by chance. Listeners judged the familiar utterances of DB to be a bit better under normal conditions. For TB the perceptual difference is increased and includes the less familiar vowels. GF produced no perceptually different imitations under nerve block except for the syllables [pi] and [peI]. However, we have no nerve block alone condition for GF, so with the nerve block, the subject could not hear himself and his palate was thickened with alginate. Under this combined condition, the addition of a nerve block was noticeable on 72% of the [i] and [eɪ] utterances.

Auditory masking, too, has only a small effect on listener judgments and affects different syllables for different speakers: DB the fricatives, TB the vowels, and GF all but the novel vowels. For the tapes of the non-nerve block speakers, we asked listeners to check the worse imitation but to mark it with an X if it were much worse. Imitations under auditory masking were judged to be much worse in [X] for speaker #1, for [Y] for speakers #2 and #4, but not much worse for any syllable of speaker #3.

Alginate placed on the alveolar ridge is more disruptive, according to perceptual judgments, than is either auditory masking or lingual nerve block. The altered vocal tract produces a more noticeable decrement, on the average, to the fricatives than to the vowels. It is especially disruptive to [ʃ] but also affects [i], while sparing (perhaps even facilitating) [ϕ] and [y].

Summary of Results

Putting the results of the various analyses together, we find that:

(1) For four speakers, familiar syllables were more noticeably affected by adverse feedback conditions than were less familiar syllables, but for three out of seven speakers, familiarity of syllables did not contribute to perceived differences among conditions. When there was an effect of familiarity, the familiar fricatives were more affected than vowels by adverse speaking conditions;

(2) Speakers made perceptually intelligible imitations of familiar syllables under all speaking conditions, although more acoustic variation was evident for [ʃ] than for the other English phones;

(3) Speakers differed in their ability to imitate non-familiar syllables with two subjects (DB and GF) making a variety of attempts to produce [X] and one subject (TB) consistently substituting [ʃ] for [X] and [ʒ] for [Y];

(4) Non-English [y] and [ϕ] were more closely approximated than were the fricatives by all three speakers, with listeners perceiving the produc—
tions as similar under all speaking conditions. Appropriate O0 activity was evident for rounded vowels and the F2 was usually lower than for their unrounded cognates as would be expected given the longer vocal tract resulting from lip rounding. In agreement with these indications of lip rounding, the less familiar vowels were judged by the phonetician to be almost foreign under normal speaking conditions;

(5) Listeners, whether expert or naive, preferred the imitations made under normal speaking conditions to those under any of the adverse speaking conditions. Imitations produced with lingual anesthesia were preferred to those produced with masking or with Alginate. Combined conditions were judged worse than any single condition;

(6) Nerve block produced a small effect on all syllable types;

(7) Auditory masking affected some syllable types more than others, depending on the speaker. In general, the F2 frequency for vowels was lower with masking;

(8) Vocal tract shape change affected [ʃ] and [i] particularly, with both spectrographic and EMG evidence of tongue retraction;

(9) There was no evidence of learning. Subjects apparently knew how to approximate [y] and [ϕ] without trials, but for [X] and [Y] they tried and failed in the time given (13 trials).

DISCUSSION

Studies of speaker compensation under difficult speaking conditions have concurred in their results, indicating that speakers are able to produce acceptable speech patterns despite bite blocks between the teeth (Lindblom, Lubker, & Gay, 1979; Fowler & Turvey, 1980) forcing a change in motor activity and despite conditions such as auditory masking and anesthesia, changing sensory information (Borden, 1979). The present study is the first, however, to manipulate the familiarity of the phonetic material. By contrasting familiar utterances with less familiar utterances, the importance of learning to motor control may be evaluated since the familiar utterances have been well practiced relative to the less familiar utterances.

The problem lies in disambiguating the perceptual and productive aspects of the control systems. To produce a skilled motor act such as a well-learned speech event, the speaker presumably makes reference to an internal representation of the utterance in the form of a perceptual target and then effects the appropriate motor coordinations according to known production rules. To produce a relatively unfamiliar motor act such as a foreign speech event, the speaker presumably refers to a more poorly formed perceptual target and enacts a motor program based on less well known production rules. Familiarity, then, influences both the perceptual target and the production.

The finding in the present study, that more familiar utterances are in some subjects more vulnerable to changes in speaking conditions than are the
less familiar utterances, lends support to the hypothesis that criteria for the perceptual targets that speakers have internalized are more detailed for familiar than for unfamiliar utterances, and that the motor programs used to produce them are more refined and well practiced. Loss of information needed to sharpen the match between speakers' actual output and intended output may result in small acoustic differences perceptible by listeners. The same loss of information about production of less familiar utterances resulted in these speakers producing less perceptually noticeable effects across speaking conditions. It follows from the same hypothesis that the less familiar utterances would be represented in the speakers' mind with a set of auditory, oro-sensory criteria that might be less well defined than for familiar utterances, as well as with a more poorly practiced set of production strategies. The use of oro-sensory information for the fine shaping of speech events has been suggested in the work of Stevens and Perkell (1977).

How is one to infer that these differences arise from differences among the productions of the speakers and not simply from the expectations of the listeners making the perceptual judgments? We know that listeners tend to categorize allophonic variations according to the phoneme systems of their own language and, further, tend to ignore small acoustic differences within their own phonemic categories (Liberman, Harris, Hoffman, & Griffith, 1957). According to the principles of categorical perception, then, English-speaking listeners would be expected to ignore differences among English-like utterances that they might notice among more foreign utterances. To the degree that the listeners in this study noticed the differences in the familiar utterances more than the unfamiliar, it can be inferred that the differences were real: they existed in the speech productions and not solely in the perceptions made by listeners.

Further support for this inference comes from the agreement with English listener judgments of an expert listener for whom the "unfamiliar" utterances were native to his language. Finally, the unfamiliar syllables (especially the consonants) were so poorly imitated and so variable, even under normal speaking conditions, that the differences between speaking conditions were relatively unimportant, whereas the imitations of the familiar syllables were remarkably good in all conditions, but small differences across speaking conditions were perceptible.

Familiarity of the phonetic material was not a significant factor for some of the speakers, indicating that for these speakers, loss of information about their own speech made no more difference in their performance, whether the performance involved well-learned or novel speech productions. The implication here is that control was essentially preplanned; with little evidence of the fine tuning of the well-learned utterances shown by three of the seven speakers.

The idea of feedforward or preprogrammed control of motor systems in speech is consistent with recent findings in the motor control literature cited in the Introduction. The compensatory motor patterns evidenced by people and by animals despite conditions that require changes in motor coordination or that remove sensory information argue for a control system with extremely rapid adaptability features. Some theorists account for the compensatory power of such motor systems by suggesting that under difficult
circumstances the motor plan is compared to the motor performance through afferent systems (Evarts, 1971; MacNeilage, 1970) or that the efferent program is simply simulated and matched with simulated afferent information, thus enabling the system to adjust by prediction without waiting for actual performance (Lindblom et al., 1979). Other theorists account for the compensations by suggesting that the equilibrium points for final positions are specified and any interference with one part of the system is adjusted for by another part of what is essentially a vibratory system (Bernstein, 1967; Kelso & Holt, 1980) or a coordinative structure (Fowler & Turvey, 1978).

There was no evidence of learning in the 13 trials produced by each subject of the less familiar fricatives [X] and [γ]. The speakers apparently failed to make use of information provided through feedback mechanisms to quickly shape a novel speech gesture. The well-programmed production rules appropriate to the speakers' language seemed in many cases to override any attempts to match a new perceptual target. It is impossible to determine from this study whether the difficulty in making the appropriate changes arises from a poor perceptual image of the target, from inadequate and poorly practiced production rules, or from a combination of perception and production factors.

There was less difficulty with [py] and [pɔ] according to listener judgments. The internal formation of some auditory-perceptual target may well have been less demanding than for the unfamiliar fricatives. If the perceptual image is easier to elicit, might it be because the production rules are not far from sounds produced in English? Although there might be a slight difference in tongue elevation and fronting for [i] and [γ], the gesture itself is not novel, nor is the gesture of lip rounding. Subjects seemed to make generally acceptable [py] and [pɔ] and the fact that they were so little affected by loss of auditory feedback indicates that the strategy taken was relatively simple (lip rounding; none of the conditions affected the lips) and may have been controlled by feedforward or open loop instructions to round the lips.

The implications of this kind of study for second language learning are obvious. We know a bit more about the ways in which perception precedes production in children learning their first language (Menyuk & Anderson, 1969; Strange & Broen, 1980; McReynolds, 1978) than we do about the ways in which perception and production may interact in adults learning a new language (Mackay, 1970; Goto, 1971; Williams, 1974; Borden, 1980).

These data suggest that for adults the basic articulation responses for speech may operate under an automatic open loop motor system, with the fine tuning of such responses resting upon the availability of a well-defined perceptual target and information on the sounds and oral sensations produced—at least for the production of continuants such as vowels and fricatives.

Future research might explore whether such feedback information can contribute to more rapid speech events as well as continuants. Also, it would be interesting to try to measure separately the development of a new perceptual target or image and the development of new motor strategies, in order to evaluate their respective contributions to the production of new speech sounds.
REFERENCES

Goodwin, G. M., & Luschei, E. S. Effects of destroying the spindle afferents from the jaw muscles upon mastication in monkeys. Journal of Neurophysiology, 1974, 37, 967-981.
Lane, H. L., & Tranel, B. The Lombard sign and the role of hearing in speech. Journal of Speech and Hearing Research, 1971, 14, 677-709.


