Effects of Training on the Acoustic–Phonetic Representation of Synthetic Speech

Alexander L. Francis
Purdue University

Howard C. Nusbaum
Kimberly Fenn
University of Chicago

Purpose: Investigate training-related changes in acoustic–phonetic representation of consonants produced by a text-to-speech (TTS) computer speech synthesizer.
Method: Forty-eight adult listeners were trained to better recognize words produced by a TTS system. Nine additional untrained participants served as controls. Before and after training, participants were tested on consonant recognition and made pairwise judgments of consonant dissimilarity for subsequent multidimensional scaling (MDS) analysis.
Results: Word recognition training significantly improved performance on consonant identification, although listeners never received specific training on phoneme recognition. Data from 31 participants showing clear evidence of learning (improvement ≥ 10 percentage points) were further investigated using MDS and analysis of confusion matrices. Results show that training altered listeners' treatment of particular acoustic cues, resulting in both increased within-class similarity and between-class distinctiveness. Some changes were consistent with current models of perceptual learning, but others were not.
Conclusion: Training caused listeners to interpret the acoustic properties of synthetic speech more like those of natural speech, in a manner consistent with a flexible-feature model of perceptual learning. Further research is necessary to refine these conclusions and to investigate their applicability to other training-related changes in intelligibility (e.g., associated with learning to better understand dysarthric speech or foreign accents).

KEY WORDS: intelligibility, synthetic speech, listener training, perceptual learning

Experience with listening to the speech of a less intelligible talker has been repeatedly shown to improve listeners' comprehension and recognition of that talker's speech, whether that speech was produced by a person with dysarthria (Hustad & Cahill, 2003; Liss, Spitzer, Caviness, & Adler, 2002; Spitzer, Liss, Caviness, & Adler, 2000; Tjaden & Liss, 1995), with hearing impairment (Boothroyd, 1985; McGarr, 1983), or with a foreign accent (Chaiklin, 1955; Gass & Varonis, 1984), or by a computer text-to-speech (TTS) system (Greenspan, Nusbaum, & Pisoni, 1988; Hustad, Kent, & Beukelman, 1998; Reynolds, Isaacs-Duvall, & Haddox, 2002; Reynolds, Isaacs-Duvall, Sheward, & Rotter, 2000; Rousenfell, Zucker, & Roberts, 1993; Schwab, Nusbaum, & Pisoni, 1985). Although experience-related changes in intelligibility are well documented, less is known about the cognitive mechanisms that underlie such improvements.

Liss and colleagues (Liss et al., 2002; Spitzer et al., 2000) have argued that improvements in the perception of dysarthric speech derive, in part,

from improvements in listeners' ability to map acoustic–phonetic features of the disordered speech onto existing mental representations of speech sounds (phonemes), similar to the arguments presented by Nusbaum, Pisoni, and colleagues regarding the learning of synthetic speech (Duffy & Pisoni, 1992; Greenspan et al., 1988; Nusbaum & Pisoni, 1985). However, although Spitzer et al. (2000) showed evidence supporting the hypothesis that familiarization-related improvements in intelligibility are related to improved phoneme recognition in ataxic dysarthric speech, their results do not extend to the level of acoustic features. Indeed, no study has yet shown a conclusive connection between word learning and improvements in the mapping between acoustic–phonetic features and words or phonemes, in either dysarthric or synthetic speech. In the present study we investigated the way that acoustic–phonetic cue processing changes as a result of successfully learning to better understand words produced by a TTS system.

TTS systems are commonly used in augmentative and alternative communication (AAC) applications. Such devices allow users with limited speech production capabilities to communicate with a wider range of interlocutors and have been shown to increase communication between users and caregivers (Romski & Sevcik, 1996; Schepis & Reid, 1995). Moreover, TTS systems have great potential for application in computerized systems for self-administered speech or language therapy (e.g., Massaro & Light, 2004).[1] Formant-based speech synthesizers such as DECtalk are among the most common TTS systems used in AAC applications because of their low cost and high versatility (Hustad et al., 1998; Koul & Hester, 2006). Speech generated by formant synthesizers is produced by rule—all speech sounds are created electronically according to principles derived from the source-filter theory of speech production (Fant, 1960). Modern formant synthesizers are generally based on the work of Dennis Klatt (Klatt, 1980; Klatt & Klatt, 1990).

One potential drawback to such applications is that speech produced by rule is known to be less intelligible than natural speech (Mirenda & Beukelman, 1987, 1990; Schmidt-Nielsen, 1995), in large part because such speech provides fewer valid acoustic phonetic cues than natural speech. Moreover, those cues that are present vary less across phonetic contexts and covary with one another more across multiple productions of the same phoneme than they would in natural speech (Nusbaum & Pisoni, 1985). Furthermore, despite this overall increased regularity of acoustic patterning compared with natural speech, in speech synthesized by rule there are often errors in synthesis such that an acoustic cue or combination of cues that were generated to specify one phonetic category actually cues the perception of a different phonetic category (Nusbaum & Pisoni, 1985). For example, the formant transitions generated in conjunction with the intended production of a [d] may in fact be more similar to those that more typically are heard to cue the perception of a [g].

Previous research has shown that training and experience with synthetic speech can significantly improve intelligibility and comprehension of both repeated and novel utterances (Hustad et al., 1998; Reynolds et al., 2000, 2002; Rousenfell et al., 1993; Schwab et al., 1985). Such learning can be obtained through the course of general experience (i.e., exposure), by listening to words or sentences produced by a particular synthesizer (Koul & Hester, 2006; Reynolds et al., 2002) as well as from explicit training (provided with feedback about classification performance or intended transcriptions of the speech) of word and/or sentence recognition (Greenspan et al., 1988; McNaughton, Fallon, Tod, Weiner, & Neisworth, 1994; Reynolds et al., 2000; Schwab et al., 1985; Venkatagiri, 1994). Thus, listeners appear to learn to perceive synthetic speech more accurately based on listening experience even without explicit feedback about their identification performance.

Research on the effects of training on consonant recognition is important from two related perspectives. First, a better understanding of the role that listener experience plays in intelligibility will facilitate the development of better TTS systems. Knowing more about how cues are learned and which cues are more easily learned will allow developers to target particular synthesizer properties with greater effectiveness for the same amount of work, in effect aiming for a voice that, even if it is not completely intelligible right out of the box, can still be learned quickly and efficiently by users and their frequent interlocutors.

More important, a better understanding of the mechanisms that underlie perceptual learning of synthetic speech will help in guiding the development of efficient and effective training methods, as well as advancing understanding of basic cognitive processes involved in speech perception. Examining the effects of successful training on listeners' mental representations of speech sounds will provide important data for

[1]Note that the Massaro and Light (2004) used speech produced by unit selection rather than formant synthesis by rule. These methods of speech generation are very different, and many of the specific issues discussed in this article may not apply to unit selection speech synthesis because these create speech by combining prerecorded natural speech samples that should, in principle, lead to improved acoustic phonetic cue patterns (see Huang, Acero, & Hon, 2001, for an overview of different synthesis methods). However, Hustad, Kent, and Beukelman (1998) found that DECtalk (a formant synthesizer) was more intelligible than MacinTalk (a diphone concatenative synthesizer). Although the diphone concatenation used by MacinTalk is yet again different from the unit selection methods used in the Festival synthesizer used by Massaro and Light (2004), Hustad et al.'s findings do suggest that concatenative synthesis still fails to provide completely natural patterns of the acoustic phonetic cues as expected by naïve listeners despite being based on samples of actual human speech.

developing more effective listener training methods, and this benefit extends beyond the domain of synthetic speech, relating to all circumstances in which listeners must listen to and understand poorly intelligible speech. Previous research has shown improvements in a variety of performance characteristics as a result of many different kinds of experience or training. Future research is clearly necessary to map out the relation between training-related variables such as the type of speech to be learned (synthetic, foreign accented, Deaf, dsyarthric), duration of training, the use of feedback, word versus sentence-level stimuli, and active versus passive listening on the one hand, and measures of performance such as intelligibility, message comprehension, and naturalness on the other. To guide the development of such studies, we argue that it would be helpful to understand better how intelligibility can improve.

To carry out informed studies about how listeners might best be trained to better understand poorly intelligible speech, it would be helpful to have a better sense of how training does improve intelligibility in cases in which it has been effective. One way to do this is by investigating the performance of individuals who have successfully learned to better understand a particular talker to determine whether the successful training has resulted in identifiable changes at a specific stage of speech understanding. In the present study, we investigated one of the earliest stages of speech processing, that of associating acoustic cues with phonetic categories.

Common models of spoken language understanding typically posit an interactive flow of information, integrating a more or less hierarchical bottom-up progression in which acoustic–phonetic features are identified in the acoustic signal and combined into phonemes, which are combined into words, which combine into phrases and sentences. This feedforward flow of information is augmented by or integrated with the top-down influence of linguistic and real-world knowledge, including statistical properties of the lexicon such as phoneme co-occurrence and sequencing probabilities, phonological and semantic neighborhood properties as well as constraints and affordances provided by morphological and syntactic structure, pragmatic and discourse patterns, and knowledge about how things behave in the world, among many other sources. In principle, improvements at any stage or combination of stages of this process could result in improvements in intelligibility, but it would be inefficient to attempt to develop a training regimen that targeted all of these stages equally. In the present article, we focus on improvements in the process of acquiring acoustic properties of the speech signal and interpreting them as meaningful cues for phoneme identification.

Researchers frequently draw on resource allocation models of perception (e.g., Lavie, 1995; Norman & Bobrow, 1975)[2] to explain the way in which poor cue instantiation in synthetic speech leads to lower intelligibility. According to this argument, inappropriate cue properties lead to increased effort and attentional demand for recognizing synthetic speech (Luce, Feustel, & Pisoni, 1983) because listeners must allocate substantial cognitive resources (attention, working memory) to low-level processing of acoustic properties at the expense of higher level processing such as word recognition and message comprehension, two of the main factors involved in assessing intelligibility (Drager & Reichle, 2001; Duffy & Pisoni, 1992; Nusbaum & Pisoni, 1985; Nusbaum & Schwab, 1986; Reynolds et al., 2002). Thus, one way that training might improve word and sentence recognition is by improving the way listeners process those acoustic–phonetic cues that are present in the signal. Training to improve intelligibility should result in learners relying more strongly on diagnostic cues (cues that reliably distinguish the target phoneme from similar phonemes) whether those cues are the same as the listener would attend to in natural speech. Similarly, successful listeners must learn to ignore, or minimize their reliance on, nondiagnostic (misleading and/or uninformative) cues, even if those cues would be diagnostic in natural speech.

To better understand how perceptual experience changes in listeners' relative weighting of acoustic cues, it is instructive to consider general theories of perceptual learning (e.g., Gibson, 1969; Goldstone, 1998). According to such theories, training should serve to increase the similarity of tokens within the same category (acquired similarity) while increasing the distinctiveness between tokens that lie in different categories (acquired distinctiveness), thereby increasing the categorical nature of perception. Speech researchers have successfully applied specific theories of general perceptual learning (Goldstone, 1994; Nosofsky, 1986) to describing this process in first- and second-language learning (Francis & Nusbaum, 2002; Iverson et al., 2003). Such changes may come about through processes of unitization and separation of dimensions of acoustic contrast as listeners learn to attend to novel acoustic properties and/or ignore familiar (but nondiagnostic) ones (Francis & Nusbaum, 2002; Goldstone, 1998), or they may result simply from changing the relative weighting of specific features (Goldstone, 1994; Iverson et al., 2003; Nosofsky, 1986).

We note, however, that although acquired similarity and distinctiveness are typically considered from the perspective of phonetic categories, such that training increases the similarity of tokens within one category and

[2] See Drager and Reichle (2001), Pichora-Fuller, Schneider, & Daneman (1995), Rabbitt (1991), and Tun and Wingfield (1994) for specific examples of the application of such models to speech perception.

increases the distinctiveness (decreases the similarity) between tokens in different categories, more sophisticated predictions are necessary when considering the effects of training on multiple categories simultaneously. Because many categories differ from one another according to some features while sharing others, a unidimensional measure of similarity is not particularly informative. For example, in natural speech the phoneme /d/ shares with /t/ those features associated with place of articulation (e.g., second formant transitions, spectral properties of the burst release), but the two differ according to those features associated with voicing. Thus, one would expect a [d] stimulus to become more similar to a [t] stimulus along acoustic dimensions correlated with place of articulation, but more different along those corresponding to voicing. For this reason, it is important to examine changes in perceptual distance along individual dimensions of contrast, not just changes in overall similarity.

In the present experiment we used multidimensional scaling (MDS) to identify the acoustic–phonetic dimensions that listeners use in recognizing the consonants of a computer speech synthesizer. By examining the distribution of stimulus tokens along these dimensions before and after successful word recognition training, we can develop a better understanding of the kinds of changes that learning can cause in the cue structure of listeners' perceptual space. There is a long history of research that uses MDS to examine speech perception using this approach. In general, much of this work reduces the perception of natural speech from a representation consisting of 40 or so American English individual phonemes to a much lower dimensional space corresponding roughly to broader classes of phonetic-like features similar to manner, place, and voicing (e.g., Shepard, 1972; Soli & Arabie, 1979; Teoh, Neuburger, & Svirsky, 2003). For natural speech, the relative spacing of sounds along these dimensions provides a measure of discriminability of phonetic segments: Sounds whose representations lie closer to one another on a given dimension are more confusable; more distant ones are more distinct. Across the whole perceptual space, the clustering of speech sound representations along specific dimensions corresponds to phonetically "natural" classes (Soli, Arabie, & Carroll, 1986). For example, members of the class of stop consonants should lie close to one another along manner-related dimensions (e.g., abruptness of onset, harmonic-to-noise ratio) because they are quite confusable according to these properties.

Poor recognition of synthetic speech (at the segmental level) is due in large part to increased confusability among phonetic segments relative to natural speech (cf. Nusbaum & Pisoni, 1985). Therefore, improved intelligibility of synthetic speech should be accompanied by increases in the relative distance among representations of sounds in perceptual space. Of course, improvements in dimensional distances would not necessarily require any changes in the structure of the space. Reducing the level of confusion between [t] and [s], for example, would not necessarily require a change in the perceived similarity of all stops relative to all fricatives, nor does it require any other kind of change that would necessarily move the structure of the perceptual space in the direction of normal phonetic organization. To take one extreme example, each phoneme could become associated with a unique (idiosyncratic) acoustic property such that all sounds become distinguished from all others along a single, unique dimension. However, this would require establishing a new dimension in phonetic space that has no relevance to the vast majority of natural speech sounds heard each day and, thus, would entail treating the phonetic classification of synthetic speech as different from all other phonetic perception. On the other hand, if perceptual learning operates to restructure the native phonetic space, it would maintain the same systematic category relations used for all speech perception (cf. Jakobson, Fant, & Halle, 1952). Indeed, most current theories of perceptual learning focus on changes to the structure of the perceptual space. Learning is understood as changing the relative weight given to entire dimensions or regions thereof (Goldstone, 1994; Nosofsky, 1986). If this is indeed the way in which perceptual learning of speech operates, then we would expect the perceptual effects of training related to improved intelligibility to operate across the phonetic space, guided by structural properties derived from the listener's native language experience. That is, we would expect that successful learning of synthetic speech should result in the development of a more natural configuration of phonetic space, in the sense that sounds should become more similar along dimensions related to shared features, and more distinct along dimensions related to contrastive features.

We should note, however, that such improvements could come about in two ways. For the most part, it is reasonable to expect that the dimensions that are most contrastive in the synthetic speech should correspond relatively well to contrastive dimensions identified for natural speech, as achieving such correspondence is a major goal of synthetic speech development. Because untrained listeners (on the pretest) will likely attend to those cues that they have learned are most effective in listening to natural speech (see Francis, Baldwin, & Nusbaum, 2000), the degree to which the synthetic speech cues correspond to those in natural speech will determine (or strongly bias) the degree of similarity between the configuration of phonemes within the acoustic–phonetic space derived from the synthetic speech and

that of natural speech. If this correspondence is good, learning should appear mainly as a kind of "fine tuning" of an already naturally structured acoustic–phonetic space. Individual stimuli should move with respect to one another, reflecting increased discriminability (decreased confusion) along contrastive dimensions and/or increased confusion along noncontrastive dimensions, but the overall structure of perceptual space should not change much: Stop consonants should be clustered together along manner-related dimensions. On the other hand, in those cases in which natural acoustic cues are not well represented within the synthetic speech, listeners' initial pattern of cue weighting (based on experience with natural cues and cue interactions) will result in a perceptual space in which tokens are not aligned as they would be in natural speech. In this case, improved intelligibility may require the adoption of new dimensions of contrast. That is, learners may show evidence of using previous unused (or underused) acoustic properties to distinguish sounds that belong to distinct categories (Francis & Nusbaum, 2002), as well as reorganizing the relative distances between tokens along existing dimensions.

Thus, two patterns of change in the structure of listeners' acoustic–phonetic space may be expected to be associated with improvements in the intelligibility of synthetic speech. First, listeners may learn to rely on new, or different, dimensions of contrast, similar to the way in which native English speakers trained on a Korean stop consonant contrast learned to use onset f0 (Francis & Nusbaum, 2002). Such a change would be manifest in terms of an increase, from pretest to posttest, in the total number of dimensions in the best fitting MDS solution (if a new dimension is added), or, at least, a change in the identity of one or more of the dimensions (cf. Livingston, Andrews, & Harnad, 1998) as listeners discard less effective dimensions in favor of better ones. In addition (or instead), listeners may also reorganize the distances between mental representations of stimuli along existing dimensions. This possibility seems more likely to occur in cases in which the cue structure of the synthetic speech is already similar to that of natural speech. This kind of reorganization would be manifest primarily in terms of an increasing similarity between representations of phonemes within a single natural class as compared with those in distinct classes, along those dimensions that are shared by members of that class. For example, we would expect the representations of stop consonants to become more similar along dimensions related to manner distinctions, even as they become more distinct along, for example, voicing-related dimensions. Thus, training should result in both improved clustering of natural classes and improved distinctiveness across classes, but which is observed for a particular set of sounds will depend on the dimensions chosen for examination.

Method

Participants

Fifty-seven young adult (ages 18–47)[3] monolingual native speakers of American English (31 women, 26 men) participated in this experiment. All reported having normal hearing with no history of speech or learning disability. All were students or staff at the University of Chicago, or residents of the surrounding community. None reported any experience listening to synthetic speech.

Stimuli

Three sets of stimuli were constructed for three kinds of tasks: consonant identification, consonant difference rating (for MDS analysis), and training (words). The stimuli for the identification task consisted of 14 CV syllables containing the vowel [a], as in *father*. The 14 consonants were [b], [d], [g], [p], [t], [k], [f], [v], [s], [z], [m], [n], [w], and [j]. The stimuli for the difference task consisted of every pairwise combination of these syllables including identical pairs (196 pairs in all) with approximately 150-ms interstimulus interval between them. The stimuli used for training consisted of a total of 1,000 phonetically balanced (PB), monosyllabic English words (Egan, 1948). The PB word lists include both extremely common (frequent, familiar) monosyllabic words such as *my, can,* and *house* as well as less frequent or less familiar words such as *shank, deuce,* and *vamp*.

Stimuli were produced with 16-bit resolution at 11025 Hz by a cascade/parallel TTS system, rsynth (Ing-Simmons, 1994, based on Klatt, 1980), and stored as separate sound files. Subsequent examination of the sound files revealed no measurable energy above 4040 Hz, suggesting that setting the sampling rate to 11025 Hz did not, in fact, alter the range of frequencies actually produced by the synthesizer. That is, the synthesizer still produced signals that would be capable of being sampled at a rate of 8000 Hz without appreciably affecting their sound. Impressionistically, the rsynth voice is quite similar to that of early versions of DecTalk. Stimuli were presented binaurally at a comfortable listening level (approximately 70 dB SPL as measured at the headphone following individual test sessions) over Sennheiser HD430 headphones.

Procedure

Participants were assigned to one of four groups. Testing was identical for all four groups, but training differed. The first ($n = 9$) and third ($n = 20$) groups received training with trial-level feedback in an active response

[3]All but 3 participants were between the ages of 18 and 25. The 3 were 32, 33, and 47, respectively.

(stimulus-response-feedback) format (henceforth, groups feedback 1 and feedback 2, respectively), the second group (n = 19) received a combination of active (but without feedback) and passive training (stimulus paired with text, with no response requested; henceforth, group no-feedback), and the fourth (control) group (n = 9) received no training at all. A control group was included because we wanted to be able to determine whether mere participation in the two sets of testing could have been sufficient to induce learning, at least to some degree.

It should be noted that, despite differences between the training supplied to the three trained groups, this study was not intended to serve as a test of training method efficacy. Rather, the differences between groups arose chronologically. After the first 18 participants had completed the study (randomly assigned to either feedback 1 or the control group), the results of another synthetic speech training study in our lab (Fenn, Nusbaum, & Margoliash, 2003) suggested that it should be possible to achieve a higher rate of successful learning (measured in terms of the number of participants achieving an increase of at least 10 percentage points in consonant recognition) with a different training method. Thus, the next 19 participants were assigned to the no-feedback condition. When this method was determined to result in no greater success rate and to have significant drawbacks for the present study including the inability to derive measures of word recognition during training that would be statistically comparable to those obtained from the first and fourth groups, the final 20 participants (feedback 2) were trained using methods as close as possible to those used for the feedback 1 group. All differences between feedback 1 and feedback 2 resulted from differences in experiment control system programming after switching from an in-house system implemented on a single Unix/Linux computer to the commercial E-Prime package (Schneider, Eschman, & Zuccolotto, 2002) that could be run on multiple machines simultaneously. Finally, the decision to assign only 9 participants to the untrained control group was based on a combination of observations: First, none of the 9 original control participants showed any evidence of learning from pretest to posttest, suggesting that including more participants in this group would be superfluous, and, second, the number of participants who failed to show significant learning despite training made it advisable to include as many participants as possible in the training condition in order to ensure sufficient results for analysis.

Results suggest that there was no difference between training methods with respect to performance on consonant recognition (see the Results section), but because this study was not intended to explore differences between training methods, no measure of word recognition was included in the testing sessions. Moreover, differences in training methods preclude direct comparison of word recognition between groups (specifically the no-feedback group versus the feedback 1 and feedback 2 groups, who received feedback on every trial). Thus, although it would be instructive to compare training method efficacy in future research, the results of the present study can only address such issues tangentially.

Testing. Testing consisted of a two-session pretest and an identical two-session posttest. The pre- and posttests consisted of a difference rating task (conducted in two identical blocks on the first and second days of testing) and an identification task (conducted on the second day of each test following the second difference rating block). The structure of the training tasks differed slightly across three groups of participants (see below).

The pre- and posttests were identical to one another, were given to all participants in the same order, and consisted of three blocks of testing over two consecutive sessions. In the first session, listeners were first familiarized with a set of 14 test syllables presented at a rate of approximately 1 syllable/s in random order. They then performed one block of 392 difference rating trials in random order. Trial presentations were self-paced, but each block typically took about 40–50 min (5–8 s per trial). Each trial presented one pair of syllables; listeners rated the degree of difference (if any) between the two sounds. There were two 392-trial difference rating blocks in both the pretest and the posttest (the first in Test Session 1, the second at the beginning of Test Session 2) totaling 784 pretest and 784 posttest ratings, four for each pair of stimuli.

Difference ratings were collected with slightly different methods for each group. For the first and fourth groups, listeners rated each pair of stimuli using a 10-cm slider control on a computer screen. Listeners were asked to set the slider to the far left if two syllables were identical and to move the slider farther to the right to indicate an increasing difference between the stimuli. The output of the slider object resulted in a score from 0 to 10, in increments of 0.1. For the no-feedback and feedback 2 groups, the difference rating was conducted using a 10-point (1–10), equal-appearing interval scale. Listeners were asked to click on the leftmost button shown on the computer screen if two syllables were identical and to choose buttons successively farther to the right to indicate an increasing difference between the stimuli.

The identification task consisted of 10 presentations of each of the 14 test syllables in random order. Listeners were asked to type in the initial consonant of each syllable they heard. An open-set response method was used to allow for the possibility that listeners might consistently mislabel specific tokens in informative

ways (e.g., writing *j* for /y/ sounds, possibly indicating a perception of greater-than-intended frication). No such consistent patterns were observed across listeners. Responses were scored as correct based on how a CV syllable beginning with that letter would be pronounced. For example, a response of *q* for the syllable [ka] was considered correct, because the only way to pronounce *q* in English is [k].

Training. Training for listeners in the feedback 1 group (n = 9) consisted of presentations of monosyllabic words produced in isolation by rsynth. For each word, listeners were asked to transcribe the word. If it did not sound like an English word, or if they did not know how to spell the word, the listeners were to type in a pattern of letters corresponding to the way the stimulus might be spelled in English. If a response did not match the spelling of the stimulus word, the correct spelling was displayed along with a spoken repetition of the stimulus. Listeners could not correct their spelling after seeing the correct response. If a response was correct, "correct response" was displayed and the stimulus was spoken again. There were four training sessions, each about 1 hr in duration, on separate days. In each training session, five PB-word lists (each 50 words in length) were presented. Thus, listeners were trained on 1,000 PB words. The order of lists and the order of words in each of the lists were randomized for each listener.

The second training group (no feedback; n = 19) participated in four sessions of five training blocks using methods similar to those described by Fenn et al. (2003). Each block of training began with a learning phase in which participants listened to individual stimuli while the orthographic form of the word appeared on the computer screen. Words (sound + orthography) appeared at 1,000-ms stimulus onset intervals. After 50 words were presented, participants were tested on those words. During the test phase a word was presented and the participant had 4 s to type an answer and press enter. If he or she did not respond in that time, or if the response was incorrect (using the same criteria as for the first group), that trial was scored as incorrect, and the next trial began (no feedback was provided to the listener). Between each block, participants were permitted to rest as long as they wished. With a total of five blocks of this kind of interleaved training and testing, participants received training on a total of 250 words per session.

The third training group (feedback 2; n = 20) was trained using a traditional training paradigm similar to that of feedback 1. On each trial, a word was presented to the participant. The participant was given 4 s to type in an answer and press *enter*. If the participant did not respond in that time, or if the response was incorrect (using the same criteria as those for the first group), that trial was marked as incorrect. After submitting an answer (or after 4 s), feedback was provided as for the Feedback 1 group: The answer was visually identified as "correct" or "incorrect," and participants heard a repetition of the stimulus along with presentation of the orthographic form of the word on the computer screen. The next trial was presented 1,000 ms after the feedback. Trials were again blocked in sets of 50 words, and there were again five blocks for each training session. Between each block, participants were permitted to rest until they chose to begin the next block. There were four training sessions in all.

It should be noted that, despite differences in training methods, all participants in the three trained groups heard exactly two presentations of each of the 1,000 words in the PB word list: One of these presentations was paired with the visual form of the word, whereas the other was not. No word ever appeared in more than one training trial.

The control group (n = 9) received no training at all because previous results have shown that this kind of control group performs identically (learns as little about synthetic speech) as one trained using natural speech rather than synthetic speech (Schwab et al., 1985). However, just as for the trained groups, control group listeners' pretest and posttest sessions were separated by about 5 days.

Results

Learning

Because word recognition was assessed differently across groups during training (feedback 1 and feedback 2 groups received feedback on every trial, whereas the no-feedback group did not), we did not compare changes in word recognition from the first to the last day of training. However, training improved overall consonant recognition, as measured by the consonant identification task on the pretest and posttest, from a mean score of 43.8% correct on the pretest to 57.6% correct on the posttest, $t(47) = 10.94$, $p < .001$.[4] The control listeners (n = 9) who did not receive any training between pretest and posttest showed no significant change in consonant identification, scoring 42% correct on both tests.

To be certain that the differences in training methods did not differentially affect the performance of the three trained groups, a mixed-factorial analysis of variance (ANOVA) with repeated measures of test (pretest vs. posttest) and a between-groups factor of group was carried out. Results showed the expected effect of test, $F(2, 45) = 159.78$, $p < .001$, but no significant effect of training group, $F(2, 45) = 0.292$, $p = .75$. However, there

[4]Because all values were close to the middle of the range from 0 to 1, no increase in statistical accuracy would be obtained by transforming the raw proportions prior to analysis, and none were performed.

was a significant interaction between test and training group, $F(2, 45) = 6.48$, $p = .003$. The interaction between test and group seems to indicate an overall greater magnitude of learning for feedback 1 (improving from 41.8% to 64.1%) over no-feedback (44.7% to 56.8%) and feedback 2 (43.9% to 55.5%) groups, possibly related to minor differences in training methods (see Table 1).[5] However, post hoc (Tukey's honestly significant difference [HSD]) analysis showed no significant difference ($p > .05$ for all comparisons) between pairs of groups on either the pretest or the posttest. Moreover, all three groups showed a significant increase in proportion correct from pretest to posttest ($p < .03$ for all three comparisons). These two findings strongly suggest that all three groups were fundamentally similar in terms of the degree to which they learned.

To investigate the effects of successful training on the structure of perceptual space, listeners from the three trained groups were reclassified according to their improvement in consonant recognition. Those listeners who showed an overall improvement in identification of at least 10 percentage points on the consonant identification task were classified as the "strong learner" group, regardless of training group membership. Thirty-one of the original 48 participants in the training groups reached this criterion of performance.[6] The other 17 listeners (1 from feedback 1, 8 from no feedback, and 8 from feedback 2), many of whom showed modest improvement, were classified as "weak learners" to distinguish them from the 9 "untrained controls" in the control group. Scores for these groups are shown in Table 1. A two-way mixed-factorial ANOVA with one between-groups factor (strong learners vs. weak learners) and one within-group factor (pretest vs. posttest) showed a significant effect of test, $F(1, 46) = 167.06$, $p < .001$, and of learning group, $F(1, 46) = 6.13$, $p = .02$, and a significant interaction between the two, $F(1, 46) = 50.56$, $p < .001$. Post hoc (Tukey's HSD) analysis showed a significant effect of training for the weak learner subgroup, who improved from 43.3% to 48.7% correct (mean improvement of 5.4%, $SD = 4.2$), as well as for the successful learners who improved from 44.1% to 62.6% correct (mean improvement of 18.5%, $SD = 6.9$). Finally, although there was no significant difference between the strong and weak learners on the pretest (43.3% vs. 44.1%), there was one on the posttest (48.7% vs. 62.6%), demonstrating a significant difference in the degree of learning among trainees who reached criterion and those who did not.

Table 1. Pretest and posttest scores (proportion of consonants identified correctly) for the four training groups.

Group	Pretest % C	Posttest % C
Feedback 1 ($n = 9$)	41.8	64.1
No feedback ($n = 19$)	44.7	56.8
Feedback 2 ($n = 20$)	43.9	55.5
Control ($n = 9$)	42.0	42.0
Learners ($n = 31$)	44.1	62.6
Trained nonlearners ($n = 17$)	43.3	48.7

Note. "% C" indicates percentage of consonants identified correctly on each test. n = number of participants in each group.

Multidimensional Scaling

Data analysis. Analysis of listeners' perceptual spaces was carried out using multidimensional scaling (MDS). MDS uses a gradient descent algorithm to derive successively better fitting spatial representations of data reflecting some kind of proximity (e.g., perceptual similarity) within a set of stimuli. The input to the analysis is a matrix or matrices of proximity data (e.g., explicit similarity judgments), and the output consists of a relatively small set of "spatial" dimensions that can account for the distances (or perceptual similarity) of the stimuli and a set of weights for each stimulus on each of these dimensions that best accounts for the total set of input proximities given constraints provided by the researcher (e.g., number of dimensions to fit).

To compute different MDS solutions, we generated a 14 × 14 matrix of difference ratings for each participant from the difference ratings for each test (pretest and posttest), such that each cell in the matrix contained the average of that listener's four ratings of the relative degree of difference between the two consonants in that pair of stimuli, ranging from 0 to 9, on that test.[7] The individual matrices for all participants in a given group (strong learners, weak learners, and untrained controls) were then averaged across participants within each test, resulting in two matrices, one for the pretest and one for the posttest, expressing the grand average dissimilarity ratings for each pairwise comparison of consonants in the stimulus set on each test. Each of these two matrices was then submitted for separate nonmetric (monotone) MDS analysis as implemented in Statistica 6.1 (Statsoft,

[5] Note, however, that the training methods for the two feedback groups were extremely similar, whereas those for the no-feedback group differed somewhat from the other two.

[6] Note that this means that 35% of trained participants did not reach the 10-percentage-point improvement criterion. Although the question of what prevents some listeners from successfully learning a particular voice under a particular training regimen may be interesting, and potentially clinically relevant, the present study was not designed to investigate this question. Rather, the present study is focused on how listeners' mental representation of the acoustic–phonetic structure of synthetic speech changes as a result of successful learning. Thus, these participants were in effect screened out in the same way that a study on specific language impairment might screen out participants who show no evidence of language deficit.

[7] Responses for participants from Groups 1 and 4, made originally using a slider on a scale of 0 to 10, were converted to a scale from 0 to 9 by linear transformation: $d_{new} = 9 \times (d_{old}/10)$.

2003). Separate starting configurations for each matrix were computed automatically using the Guttman–Lingoes method based on a principal components analysis of the input matrix as implemented automatically in the Statistica MDS procedure.

After obtaining a solution for each test, distances among tokens were calculated automatically using a Euclidean distance metric. In addition, we calculated two kinds of ratios from the interpoint distances in the final MDS spaces. In the general case, *structure ratios* (Cohen & Segalowitz, 1990) serve as a measure of the degree of similarity within a category (e.g., different exemplars of a single phoneme) relative to that across categories (e.g., exemplars of different phonemes). A small structure ratio reflects tightly grouped categories widely separated from one another. Thus, to the extent that training-related increases in improved intelligibility derive from improved categorization of phonemes, one would expect to see a decrease in structure ratio corresponding to an increase in intelligibility. However, because all instances of a given utterance produced by a TTS system are acoustically identical, in the present experiment it was not possible to compare changes in structure ratios based on phonetic category because there was only one instance of each phoneme in the test set. This is necessarily the case for synthetic speech produced by rule: All productions of a given phoneme are either acoustically identical or differ in terms of more than their consonantal properties (e.g., preceding a different vowel). Therefore, structure ratios were calculated for three natural phonetic manner classes rather than for individual phonemes. These classes were as follows: continuants: [w], [y], [m], and [n]; stops: [b], [d], [g], [p], [t], and [k]; and fricatives: [f], [s], [v], and [z].

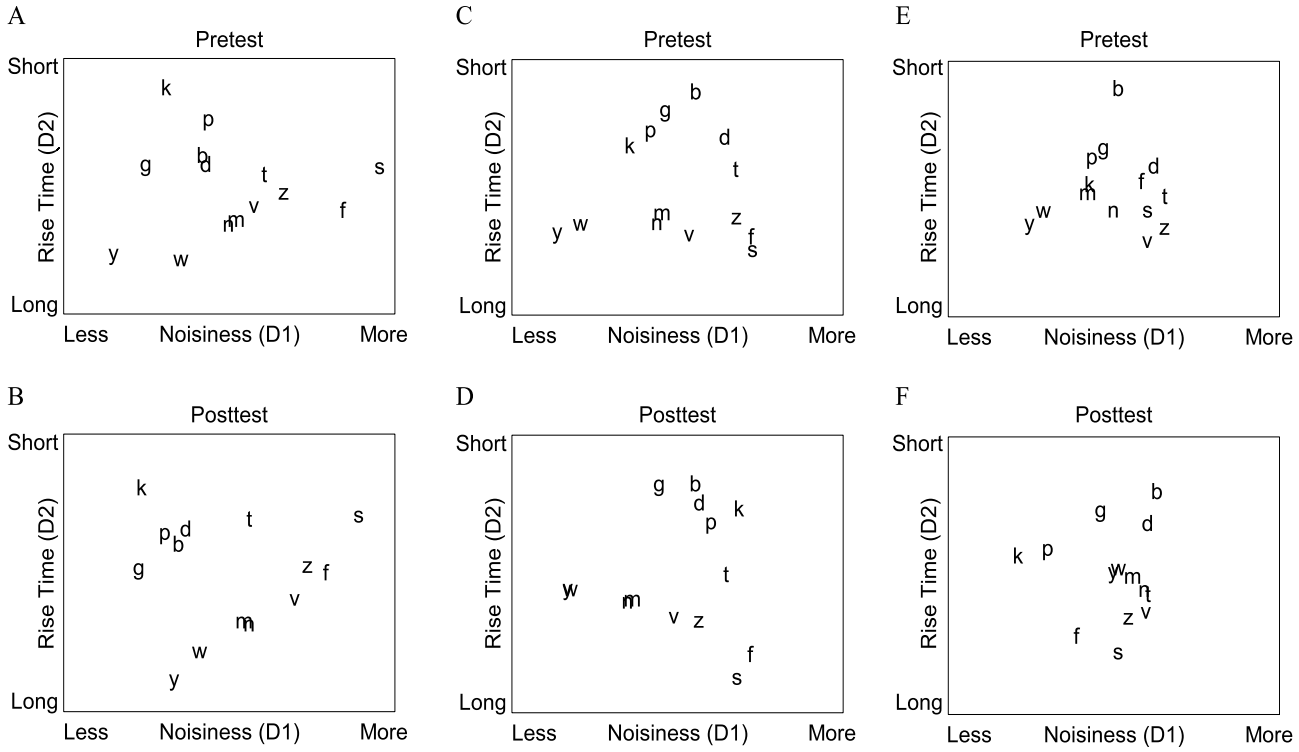
The interpretation of structure ratios based on natural classes is not as straightforward as it would be for phoneme-based structure ratios because, although it is reasonable to expect that learning should cause all exemplars of a given phoneme to become more similar and all exemplars of different phonemes to become more distinct, the same does not hold true for multiphoneme classes. For example, although two instances of [b] would presumably be perceived as more similar in every way after training, an instance of [b] and an instance of [d] should only become more similar along dimensions that do not distinguish them (e.g., those related to voicing) and should in fact become more distinct along other dimensions (e.g., those related to place of articulation). Thus, it is possible that structure ratios based on natural classes should decrease as a consequence of training for the same reasons that those based on phonemes would, but the degree to which this might occur must depend on the combination of the changes in distances between each pair of members of a given class along each individual dimension that defines the phonetic space.

However, it is possible to apply the same principles to construct structure ratios based on distances between tokens along a single dimension (e.g., one related to voicing), rather than those based on overall Euclidean distance, comparing the distance along one dimension (e.g., voice onset time [VOT]) between tokens within a given class (e.g., voiced phonemes) to differences along the same dimension between tokens in different classes (e.g., voiced vs. voiceless). In the case of such *dimension-specific structure ratios,* the predictions are identical to those for phoneme-based structure ratios. Stimuli that share a phonetic feature should become more similar along dimensions cuing that feature, whereas stimuli that possess different values of that feature should become more dissimilar. For example, one would expect to see a decrease in the VOT-specific structure ratio for the set of stimuli [b], [d], and [g] as compared with [p], [t], and [k]—the pairwise distances along the VOT dimension within the sets ([b], [d], and [g]) and ([p], [t], and [k]) should decrease relative to the pairwise distances between tokens from different sets.

In addition to dimension-specific structure ratios, *dimensional contribution ratios* were used to compare distances between individual tokens. These were calculated as the ratio of the distance between two tokens along a single dimension to the distance between those two tokens in the complete MDS solution space. Structure ratios (both overall and dimension specific) and dimension contribution ratios are invariant under linear transformation of all dimensions (i.e., dilations/contractions and reflections), and, therefore, are legitimately comparable across separately normalized solutions.

Characteristics of listeners' perceptual space. Based on the results of previous research (e.g., Soli & Arabie, 1979), four-dimensional (4D) solutions were obtained for both the pretest and posttest dissimilarity ratings matrices for each class of listener (strong learners, weak learners, and untrained controls), as shown in Figure 1. One way to measure the adequacy of an MDS solution is to compare the set of pairwise distances between tokens in the derived space to the same values in the original data set. There are a variety of such statistics, but the most commonly used one is Kruskal's Type I stress (hereafter referred to simply as *stress;* Kruskal & Wish, 1978). Stress is based on the calculation of a normalized residual sum of squares. Values below 0.1 are typically considered excellent, and those below 0.01 may indicate a degenerate solution (i.e., too many dimensions for the number of stimuli). All else being equal, stress can always be reduced by increasing the number of dimensions in a solution. However, solutions with more dimensions are typically more difficult to interpret, and interpretability of the solution is another consideration (arguably more important than stress) when deciding on the number of dimensions to use in an MDS solution

Figure 1. Dimension 1 versus Dimension 2 of the four-dimensional multidimensional scaling solution for trained strong learners' pretest (A) and posttest (B), weak learners' pretest (C) and posttest (D), and untrained control listeners' pretest (E) and posttest (F) difference ratings of synthesized consonants. Note that in the posttest figures for the two trained groups (B and D), the locations of [n] and [m] overlap almost completely, as do [t] and [n] in the untrained groups' posttest figure (F), and thus may be difficult to distinguish visually.



(see Borg & Groenen, 1997; Kruskal & Wish, 1978, for details and discussion). In the present case, for the strong learner group, stress was 0.038 for the pretest solution and 0.041 for the posttest, suggesting a good fit between the solution configuration and the original measured dissimilarity values. This was also the lowest dimensionality for which stress was under 0.1 for both the pretest and the posttest solutions, suggesting an optimal compromise between minimizing stress and number of dimensions. Similar results were obtained for the weak learners (pretest = 0.057; posttest = 0.032) and the untrained controls (pretest = 0.062; posttest = 0.061).

Figure 1 shows the first two dimensions of the 4D solution space for the strong learners', weak learners', and control listeners' pretest and posttest dissimilarity matrices.[8] In the present article we are not concerned with the specific identity of particular dimensions, and the stimulus sets were not designed for this task. However, the relative ordering of tokens along each dimension in both the pretest and the posttest solutions suggests plausible phonetic interpretations: At the right side of Dimension 1 (D1) lie voiceless fricatives, contrasting with stops and glides on the left, suggesting that the degree of noisiness heard in a given token might best characterize this dimension (compare the observation of Samuel & Newport, 1979, that periodicity may be a basic property of speech perception). On Dimension 2 (D2), the stops [p], [k], [b], [d], and [g] lie toward the top, and the other manner categories are lower down, suggesting a distinction such as stop/nonstop or, perhaps, the abruptness of onset (rise time) of the syllable as suggested by Soli and Arabie's (1979) reference to "abruptness of onset of aperiodic noise" dimension, or Stevens's (2002) "acoustic discontinuity" feature. Similar analyses show that the third and fourth dimensions correspond relatively well with the acoustic features of duration and slope of second formant transition (correlated with place of articulation), respectively. However, we focus our discussion here on the first two dimensions because the lower dimensions tend to account for the greatest proportion of

[8] The interpoint distances in MDS solutions are invariant with respect to rotation and reflection (Borg & Groenen, 1997; Kruskal & Wish, 1978) in the sense that the algorithm's identification of a particular dimension as D1 as opposed to D2 is arbitrary, as is the orientation of any given dimension (i.e., it does not matter whether low values of D1 are plotted to the left and high values to the right, or vice versa). Thus, the dimensions shown here have been reflected and/or rotated 90° where appropriate to align them in a visually more understandable manner.

Table 2. Overall structure ratios for three natural classes of consonants.

	Learners			Weak learners			Controls		
Consonant class	**Pre**	**Post**	**Diff.**	**Pre**	**Post**	**Diff.**	**Pre**	**Post**	**Diff.**
Stops	1.08	0.94	−0.14	1.10	1.12	0.01	1.21	1.26	0.05
Fricatives	0.73	0.73	0.00	0.71	0.57	−0.14	0.75	0.56	−0.19
Continuants	0.65	0.54	−0.11	0.59	0.53	−0.06	0.62	0.61	−0.01
Average	0.82	0.73	−0.09	0.80	0.74	−0.06	0.86	0.81	−0.05

Note. Pre = pretest; Post = posttest; Diff. = difference in structure ratio from pretest to posttest.

the variance of an MDS solution and because in the current results, these most clearly correspond to linguistic phonetic properties (voicing and manner) that are contrastive for most or all of the phonemes in the stimulus set.

More important than the specific identity of the dimensions shown here is the observation that the pattern of token distribution is quite similar across both pretest and posttest solutions for all three groups.[9] Thus, it appears that listeners were consistent in the cues they used for phoneme recognition. That is, training does not appear to have induced listeners to select different acoustic cues for consonant identification after training than before but may instead have altered the relative weighting that listeners gave to specific cues.

Effects of training. *Structure ratios* for stops, fricatives, and continuants for all three listener groups (strong learners, weak learners, and controls) are given in Table 2. There was an overall pattern of decrease from pretest to posttest (shown in the difference column). This decrease was greater for strong learners than for weak learners, and marginally greater for weak learners than for controls, suggesting that successful training is related to a relative increase in within-class similarity and/or increase in between-class distinctiveness, and that even the unsuccessful learners may have experienced some improvement in the organization of their perceptual space. However, these numbers must be interpreted with caution for two reasons. First, they are based on analyses of group solution spaces, and, therefore, it is not possible to conduct hypothesis-testing statistics on them. Second, they reflect the structure of multiphoneme classes, not individual phonemes, and, therefore, are composed both of distances along which tokens would be expected to become more similar following training as well as those along which they should become more distinct. Thus, further analyses were restricted to changes along individual dimensions for which specific predictions can be made, focusing on D1 and D2 as shown in Figure 1 and Tables 3 and 4.

Strong learners' changes along D1 were characterized by a decrease in the *D1-specific structure ratio* (indicating increased within-class similarity and/or increased between-class distinctiveness) for all three classes, whereas weak learners showed decreases only for stops and continuants, and control listeners showed a decrease only for continuants. On average, both trained groups showed decreases, whereas untrained controls showed an average increase. Along D2, strong learners showed an obvious decrease in the D2-specific structure ratio for stops and a marginal one for continuants. Weak learners showed only a slight decrease for continuants, whereas control listeners showed a marked decrease for all three classes. These results suggest that learning was most clearly associated with an increased perception of similarity of phonemes with similar D1 values and/or increased perception of dissimilarity between phonemes differing according to D1. This pattern was stronger for successful learners than for less successful ones and was even less apparent for untrained listeners. In contrast, learning-related changes along D2 primarily involved a relative increase in distinctiveness within the class of fricatives, and again, this pattern was stronger for the successful learners than for the less successful ones.

Patterns of change in location for specific speech sounds are also informative. As shown for strong learners in Figures 1A and 1B, before training [t] lay to the right of (at a higher value than) [v] along the first dimension, suggesting that, before training, [t] was perceived to be more fricativelike—at least, according to the acoustic properties represented by D1—than was [v]. After training the order was reversed; [v] was perceived to be more fricativelike according to D1 than [t]. This observation was supported by an analysis of the *dimensional contribution* of D1 (Table 5) to the average distance between [t] and fricatives ([f], [s], [v], and [z]), which increased from 0.46 to 0.60, suggesting that learners gave more weight to D1 when comparing [t] and fricatives on the posttest than they did on the pretest. Similarly, the contribution of D1 to the average distance

[9] The comparative noisiness of the solutions for the control listeners is likely the result of the small number of participants compared with the other two groups.

Table 3. Dimension-specific structure ratios for Dimension 1 (D1; noisiness).

	Learners (D1)			Weak learners (D1)			Controls (D1)		
Consonant class	Pre	Post	Diff.	Pre	Post	Diff.	Pre	Post	Diff.
Stops	0.57	0.48	−0.09	0.79	0.48	−0.31	0.83	1.43	0.60
Fricatives	0.60	0.28	−0.32	0.39	0.60	0.21	0.20	1.00	0.80
Continuants	0.78	0.57	−0.21	0.67	0.38	−0.29	0.72	0.51	−0.21
Average	0.65	0.44	−0.21	0.62	0.49	−0.13	0.58	0.98	0.40

between [v] and other fricatives decreased from 0.65 to 0.30, suggesting that training encouraged successful learners to give less weight to D1 when making comparisons between [v] and other fricatives. Interestingly, the contribution of D1 to the distance between [t] and other stops ([p], [k], [b], [d], and [g]) increased from 0.59 to 0.74, as did that between [v] and stops, from 0.39 to 0.72. This suggests that strong learners gave more weight to D1 when comparing both [t] and [v] to stops.

Results of the MDS analysis are augmented by an analysis of confusions (see the Appendix for raw confusion matrices). On the pretest, strong learners identified the [t] token as a fricative 143 times out of a total of 200 misidentifications (72%).[10] On the posttest, this proportion dropped to 66 out of 109 misidentifications (61%). Simultaneously, confusions of [t] with other (non-[t]) stop consonants increased from 29% (57/200) to 39% (43/109). On the pretest, [t] was misidentified as a fricative (143 times) more often than it was identified as a stop of any kind (136 times, of which only 79 were correct [t] responses). On the posttest, this ratio changed noticeably to 66 fricative responses versus 223 stop responses (including 180 correct [t] responses).

By contrast, on both the pretest and the posttest, only 2% of the strong learners' misidentifications of [v] were as a stop consonant. On the pretest, [v] was misidentified once each as [b], [d], [g], [t], and [k], whereas on the posttest, [v] was only misidentified twice, as [g]. Similarly, on the pretest, only 4% of the misidentifications of [v] were as a continuant (once each as [m] and [w], twice each as [n] and [y]); on the posttest, such confusions occurred only 6% of the time (twice as [m], four times as [n], and once as [w]). On both the pretest and the posttest, 93% of all misidentifications of [v] were as some other fricative consonant (on the pretest: as [f], 33/167 [20%], as [s], 10/167 [6%], and as [z], 113/167 [68%]; on the posttest: as [f], 9/121 [7%], as [s], 3/121 [2%], and as [z], 100/121 [83%]). The lack of major change in [v] confusions suggests that primary source of the crossover observed in Figure 1 derives from strong learners' changing mental representation of the [t] token, from relatively fricativelike to considerably less fricativelike (though only slightly more stoplike).

Discussion

Learning to understand words produced by a computer speech synthesizer improved consonant identification accuracy, although listeners were trained only on word identification. Although the improvement was comparatively modest, such improvement suggests that at least part of the training-related improvement in word intelligibility of synthetic speech was due to improvements in the way listeners processed individual phonemes, consistent with the results discussed by Spitzer et al. (2000) for dysarthric speech. In support of this conclusion, the results of the MDS analysis and analysis of confusions in explicit consonant identification showed that learning restructured listeners' phonetic space. This suggests that listeners were learning to better identify those acoustic properties of the speech that were consistent with the phonological features of the intended phoneme. Learners shifted the relative weight given to available acoustic cues such that, after training, the synthetic speech sounds were categorized in a manner more closely resembling that of natural speech: Stop consonants were more likely to be confused with other stops, fricatives with other fricatives, and so on. Thus, even a short period of training is capable of improving listeners' use of nonstandard acoustic cues, supporting the hypothesis that listener training can be a useful tool for improving segmental intelligibility under certain circumstances.

It should also be noted that changes in perceptual space were not completely restricted to the strong learner group, although in general their overall magnitude did appear to be smaller among weaker learners. For example, among the weak learners, [m] and [n] move closer to [y] and [w] after training, as reflected in the decrease in the structure ratio for continuants for this group. On the basis of the present results, it is not clear whether such a change reflects beneficial learning (i.e., learning that could have increased intelligibility) or whether it reflects irrelevant or detrimental learning. Because these

[10] Because the incidence of specific patterns of confusion was in some cases extremely small (less than one such confusion, on average, per listener), these numbers are results for the entire group as a whole, and inferential statistical analyses were not carried out because the magnitude of change is so low given the sparse samples.

Table 4. Dimension-specific structure ratios for Dimension 2 (D2; abruptness of onset).

	Learners (D2)			Weak learners (D2)			Controls (D2)		
Consonant class	**Pre**	**Post**	**Diff.**	**Pre**	**Post**	**Diff.**	**Pre**	**Post**	**Diff.**
Stops	0.53	0.40	−0.13	0.34	0.36	0.02	0.79	0.66	−0.12
Fricatives	0.50	0.70	0.19	0.28	0.41	0.13	0.80	0.32	−0.47
Continuants	0.35	0.34	−0.01	0.18	0.13	−0.05	0.41	0.24	−0.17
Average	0.46	0.48	0.02	0.27	0.30	0.03	0.66	0.41	−0.26

changes were observed in the weak learner group, we can conclude that the total effect of all such changes was significantly less useful, in terms of improving intelligibility, than was that of the changes observed in the successful learners. However, from the present results, we cannot tell whether a specific change such as decreasing the structure ratio for continuants was in itself irrelevant or detrimental, or whether it was generally beneficial but simply failed to contribute enough to improving intelligibility to affect the overall pattern of performance.

Similarly, although the increased weighting of D1 for both [t] and [v] in comparisons involving stop consonants is quite intuitive for [v] (which should be expected to differ noticeably from stops along a rise-time/abruptness-of-onset type dimension such as D1), it seems less clear why strong learners should exhibit such a weighting change in the case of [t]. That is, it is difficult to determine why training encouraged strong learners to give more weight to D1 properties of the signal when distinguishing [t] from other stops. It is possible that this change is the result of an increase in the weight given to the entire D1 dimension, as would be predicted by a model such as Nosofsky's generalized context model that operates by uniformly increasing or decreasing the weight given to individual dimensions. In this case, any detrimental effect of the increased weighting of D1 for comparisons of [t] with other stops would be outweighed by the benefits of increasing the distinctiveness of all stops (including [t]) from nonstops along this dimension. On the other hand, it is also possible that this increased weighting of D1 even for interstop comparisons could reflect a positive adaptation, perhaps, for example, contributing to improved separation of stops differing according to place of articulation. In either case, the observation that both weak learners and untrained controls show a decrease in the weight given to D1 for interstop comparisons suggests that such a decrease was not, in itself, a sufficient contributor to overall improvement in intelligibility.

In general, these results suggest that training may induce not only beneficial changes in cue weighting but also changes that are either irrelevant to improvements in intelligibility or at least that do not contribute to improvements in intelligibility to a noticeable degree. Further research is clearly necessary to explore factors that may affect the efficacy of training in specific contexts and with specific kinds of nonstandard speech.

Interestingly, the lack of a difference between the no-feedback group (who received no feedback during training) and the feedback 1 and feedback 2 groups (who did) is only partly consistent with results presented by McCandliss, Fiez, Protopapas, Conway, and McClelland (2002). They found that active training (word recognition, similar to that used here) without feedback was sufficient to improve native Japanese listeners' performance on a trained *rock–lock* (or *road–load*) continuum to a similar degree as was training with feedback, but training without feedback was not as good as training with feedback in inducing generalization to the other (untrained) continuum. In the present case, training with feedback (feedback 1 and feedback 2 groups) and without feedback (no feedback group) resulted in comparable improvement in consonant recognition performance for stimuli that were not explicitly trained (although some

Table 5. Dimensional contribution of D1 (noisiness) to the distance between specific tokens ([t] and [v]) and all (other) tokens in two natural classes (stops and fricatives).

	Learners			Weak learners			Controls		
Token/class	**Pre**	**Post**	**Diff.**	**Pre**	**Post**	**Diff.**	**Pre**	**Post**	**Diff.**
[t] / fricatives	0.46	0.60	0.14	0.17	0.32	0.15	0.13	0.38	0.24
[v] / fricatives	0.65	0.30	−0.35	0.57	0.60	0.03	0.15	0.49	0.34
[t] / stops	0.59	0.74	0.16	0.47	0.24	−0.23	0.49	0.43	−0.07
[v] / stops	0.39	0.72	0.33	0.25	0.27	0.02	0.27	0.34	0.07

of the test syllables did appear in training words; e.g., [ba] in *box* or *bomb*), suggesting that learning a different cue structure for a specific (synthetic) speaker of one's native language may be in some way less difficult, or more easily generalized, than learning the unfamiliar cue structure relevant to perception of a foreign phonetic contrast. Overall, however, the changes observed here in strong (and, to some extent, weak) learners' perceptual space also provide some insight into the mechanisms underlying phonetic learning more generally, as discussed in the following section.

Theoretical Considerations

The decrease in the D1-specific structure ratio for all classes of sounds (in the strong learner group) suggests that this property is more important as a cue for distinguishing between these classes than it is for differentiating between categories within a given class (e.g., stops). On the other hand, the increase in the D2 structure ratio for fricatives suggests that this property is more important for distinguishing between members of the class of fricatives than it is for differentiating fricatives from other classes of sounds, which is supported by the observation that a reduction in the number of confusions between [v] and [f] corresponds to an increase in the perceptual distance between these two tokens along this dimension.

In broad terms, the observed changes in listeners' perceptual (phonetic) space are consistent with localized warping (prototype magnet) models of category learning (e.g., Francis & Nusbaum, 2002; Goldstone, 1994; Guenther, Husain, Cohen, & Shinn-Cunningham, 1999; Kuhl & Iverson, 1995; see Gibson, 1969, for the theoretical basis for this kind of model). According to such models, category learning involves both acquired similarity within categories and acquired distinctiveness between categories. Such changes in perceptual space have been operationalized in terms of shifting the focus of attention given to specific dimensions, such that increasing the attention given to a particular featural dimension "stretches" it, whereas withdrawing attention "shrinks" it (e.g., Goldstone, 1994; Nosofsky, 1986). Recently, a number of researchers have incorporated this kind of "attention to dimension" (A2D) mechanism into models of speech perception and perceptual learning of novel phonetic categories by infants and second-language learners (Guenther et al., 1999; Iverson & Kuhl, 1995, 2000; Iverson et al., 2003; see Francis & Nusbaum, 2002, for discussion). The results presented here are also compatible with such models, even though the learning investigated here occurred in adults learning to understand a novel talker, not a new linguistic system. It seems plausible that this kind of model may be generally applicable for representing all kinds of low-level changes in phonetic processing (i.e., those not involving acquisition of new categories), including not only those resulting from experience with synthetic speech but, potentially, also those related to learning the speech of other highly atypical talkers such as those with dysarthria, hearing impairment, or a foreign accent (Boothroyd, 1985; Liss, Spitzer, Caviness, & Adler, 2002; Chaiklin, 1955; Gass & Varonis, 1984; McGarr, 1983; Spitzer et al., 2000).

However, in addition to these changes, we also observed local changes in the rank order of tokens along a dimension of contrast that went beyond the typical operations of stretching and shrinking described by A2D models. A2D models predict that such changes should come about only through the warping of dimensions, either as a whole (Nosofsky, 1986) or in localized areas corresponding to category boundaries (stretching) and category prototypes (shrinking; Goldstone, 1994; Guenther et al., 1999; Iverson & Kuhl, 1995). In all such models, changing the weight given to a dimension of contrast should only serve to change the relative distance between tokens—it should not change their rank order along any dimension, as was observed here for the [t] and [v] tokens. Although the present results must be interpreted with caution due to the wide range of possible factors that may influence the appearance of an MDS solution, the observed changes suggest that there may be other kinds of operations in perceptual learning than are predicted by existing dimensional warping models.

For example, it is possible that the appearance of crossovers is a consequence of the flexibility of the dimensions that structure phonetic space. Research by Schyns and colleagues (Schyns, 1998; Schyns, Goldstone, & Thibaut, 1998; Schyns & Oliva, 1997; Schyns & Rodet, 1997) suggests that perceptual features are flexible, and research on trading relations (Pisoni & Luce, 1987; Repp, 1982) and cue weighting (Mayo, Scobbie, Hewlett, & Waters, 2003) suggest that this may be particularly true for speech. In the case of the present experiment, it is possible that the D1 of the pretest is not precisely the same, in terms of the acoustic properties it reflects, as the D1 of the posttest. According to a flexible feature model, the two apparently equivalent dimensions encode the same linguistic distinctions but do so on the basis of different combinations of acoustic properties. In this case, the appearance of crossovers results directly from changes to the structure of listeners' perceptual space but not in a manner amenable to traditional A2D models.

According to a flexible feature model of phonetic learning, listeners learn to incorporate different combinations and proportions of low-level acoustic properties of the speech signal into higher level, phonetically meaningful dimensions depending on the demands of

the listening task. The appearance of crossovers results from listeners learning to modify the subcomponents of integrated sets of acoustic features that together function as complex, linguistically meaningful dimensions. As a consequence of training, listeners learned to reorganize these acoustic properties, weighting some more heavily and others less so, but again incorporating them into a functionally equivalent (but perceptually slightly different) phonetic dimension with a more appropriate rank ordering of tokens along it (see also Livingston et al., 1998; Schyns & Oliva, 1997). Such a possibility is purely speculative in the present case, and MDS analyses (alone) would not be able to evaluate such a hypothesis because of the uncertainty of dimensional identification associated with this procedure. However, there is a growing literature supporting the flexibility of visual category systems (e.g., Quinn, Schyns, & Goldstone, 2006), and it may be productive to follow up this issue in the auditory/speech domain as well.

In summary, this study demonstrated that even strict A2D models of phonetic learning can account for most of the changes observed in perceptual learning of synthetic speech. However, the present results also suggest that perceptual learning may affect the composition of the set of acoustic features that listeners attend to in developing and maintaining mental representations of speech sound categories. Further research is necessary to determine the nature of the relation between individual features and the effects of learning on these relations.

The application of a flexible feature model to the perceptual learning of unfamiliar speech sounds also allows for the possibility that listeners might learn to attend to novel acoustic cues or (also) to ignore familiar ones. Although in the present experiment we observed no evidence of a change from pretest to posttest in the number of dimensions used by listeners or in the general acoustic–phonetic properties represented by each dimension in the present study, these kinds of changes have been observed in perceptual learning in other modalities (Livingston et al., 1998) and in phonetic learning of nonnative consonant contrasts (Francis & Nusbaum, 2002). Both of these cases involved the acquisition of novel categories (not just new patterns of cues for old categories), so it is not yet known whether such effects would be observed as a consequence of learning to better understand either a TTS system or a disordered talker producing speech in the listener's native language.

In principle, the development or discovery of new acoustic–phonetic features may be more likely to occur in listeners' adaptation to dysarthric speech or the speech of individuals with anomalous vocal tracts than in learning to better understand synthetic speech, assuming that the disordered speech to be learned was still sufficiently systematic across instances of the same utterance. There are two issues here: the consistency of the individual speaker, and the relation between the speaker's cue structure and that of the listener's native system. On the one hand, inconsistent mapping between cues and linguistic units might make learning of any kind difficult or impossible (cf. Shiffrin & Schneider, 1977). Thus, little or no learning might be expected in cases in which the speaker is not consistent in his or her use of cues across instances of the same utterance (e.g., as might occur with spasmodic dysphonia or some kinds of dysarthria), and without learning one would not expect to see large, linguistically motivated changes in the structure of the listeners' acoustic–phonetic representations of that talker's speech (cf. the untrained controls in the present study). On the other hand, if learning is indeed successful, the degree to which it might involve the direction of attention to an entirely new cue may be quite high in cases in which the talker's system for speech production differs significantly from normal.

The acoustic–phonetic cue structure of synthetic speech is modeled on that of fluent, natural speech, and, therefore, the kinds of cues found in synthetic speech are presumably intentionally designed to match as closely as possible the expectations of listeners experienced with that kind of speech. That is, the acoustic features provided in synthetic speech are modeled on those found in fluent natural speech, and, therefore, approximate them to some arbitrary degree. In contrast, the diagnostic acoustic properties found in disordered speech may be (in principle) radically different from those in fluent speech (Ansel & Kent, 1992; Kent, Weismer, Kent, & Rosenbek, 1989; although see Tjaden, Rivera, Wilding, & Turner, 2005, for some properties that do not appear to differ), and the speech of individuals with anomalous vocal tracts (e.g., resulting from congenital craniofacial anomalies, or glossectomee and other oral surgeries) may even contain diagnostic acoustic properties that would be impossible in the speech of normal talkers. Further research is clearly necessary to identify acoustic–phonetic cues used in the perception of synthetic, dysarthric, and other anomalous speech, and to further explore learning-related changes in the precise composition and relative weighting of these cues.

Acknowledgments

Some of the data in this article derive from part of a doctoral dissertation submitted by the first author to the Department of Psychology and the Department of Linguistics at the University of Chicago. Some of these results were presented at the 136th meeting of the Acoustical Society of America in Norfolk, Virginia, on October 15, 1998. This work was supported, in part, by a grant from the Division of Social Sciences at the University of Chicago to the second author and by National Institutes of Health Grant R03 DC006811 (awarded to the first author). We are grateful to Lisa Goffman

and Jessica Huber for helpful comments on previous versions of this article.

This article is dedicated to the memory of Nick Ing-Simmons, author of *rsynth*.

References

Ansel, B. M., & Kent, R. D. (1992). Acoustic–phonetic contrasts and intelligibility in the dysarthria associated with mixed cerebral palsy. *Journal of Speech and Hearing Research, 35,* 296–308.

Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling: Theory and applications*. New York: Springer.

Boothroyd, A. (1985). Evaluation of speech production of the hearing impaired: Some benefits of forced-choiced testing. *Journal of Speech and Hearing Research, 28,* 185–196.

Chaiklin, J. B. (1955). Native American listeners' adaptation in understanding speakers with foreign dialect. *Journal of Speech and Hearing Disorders, 20,* 165–170.

Cohen, H., & Segalowitz, N. (1990). Cerebral hemispheric involvement in the acquisition of new phonetic categories. *Brain and Language, 38,* 398–409.

Drager, K. D. R., & Reichle, J. E. (2001). Effects of age and divided attention on listeners' comprehension of synthesized speech. *Augmentative and Alternative Communication, 17,* 109–119.

Duffy, S. A., & Pisoni, D. B. (1992). Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Language and Speech, 35,* 351–389.

Egan, J. P. (1948). Articulation testing methods. *The Laryngoscope, 58,* 955–991.

Fant, G. (1960). *Acoustic theory of speech production*. The Hague, the Netherlands: Mouton.

Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003, October 9). Consolidation during sleep of perceptual learning of spoken language. *Nature, 425,* 571–572.

Francis, A. L., Baldwin, K., & Nusbaum, H. C. (2000). Effects of training on attention to acoustic cues. *Perception & Psychophysics, 62,* 1668–1680.

Francis, A. L., & Nusbaum, H. C. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 349–366.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning, 34,* 65–89.

Gibson, E. J. (1969). *Principles of perceptual learning and development*. New York: Appleton-Century-Crofts.

Goldstone, R. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General, 123,* 178–200.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology, 49,* 585–612.

Greenspan, S. L., Nusbaum, H. C., & Pisoni, D. B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 421–433.

Guenther, F. H., Husain, F. T., Cohen, M. A., & Shinn-Cunningham, B. G. (1999). Effects of categorization and discrimination training on auditory perceptual space. *The Journal of the Acoustical Society of America, 106,* 2900–2912.

Huang, X., Acero, A., & Hon, H.-W. (2001). *Spoken language processing*. Upper Saddle River, NJ: Prentice Hall.

Hustad, K. C., & Cahill, M. A. (2003). Effects of presentation mode and repeated familiarization on intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology, 12,* 198–208.

Hustad, K. C., Kent, R. D., & Beukelman, D. R. (1998). DECtalk and MacinTalk speech synthesizers: Intelligibility differences for three listener groups. *Journal of Speech, Language, and Hearing Research, 41,* 744–752.

Ing-Simmons, N. (1994). *rsynth 2.0*. Retrieved June 19, 2006, from http://www.speech.cs.cmu.edu/comp.speech/Section5/Synth/rsynth.html

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *The Journal of the Acoustical Society of America, 97,* 553–562.

Iverson, P., & Kuhl, P. K. (2000). Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from a common mechanism? *Perception & Psychophysics, 62,* 874–886.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition, 87,* B47–B57.

Jakobson, R. C., Fant, G. M., & Halle, M. (1952). *Preliminaries to speech analysis: The distinctive features and their correlates* (Tech. Rep. No. 13). Cambridge, MA: MIT Acoustics Laboratory.

Kent, R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders, 54,* 482–499.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America, 67,* 971–995.

Klatt, D., & Klatt, L. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America, 87,* 820–857.

Koul, R., & Hester, K. (2006). Effects of repeated listening experiences on the recognition of synthetic speech by individuals with severe intellectual disabilities. *Journal of Speech, Language, and Hearing Research, 49,* 47–57.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-011). London: Sage.

Kuhl, P., & Iverson, P. (1995). Linguistic experience and the "perceptual magnet effect." In W. Strange (Ed.), *Speech perception and linguistic experience* (pp. 121–154). Baltimore: York Press.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 451–468.

Liss, J. M., Spitzer, S. M., Caviness, J. N., & Adler, C. (2002). The effects of familiarization on intelligibility and lexical segmentation in hypokenetic and ataxic dysarthria. *The Journal of the Acoustical Society of America, 112,* 3022–3030.

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 732–753.

Luce, P. A., Feustel, T. C., & Pisoni, D. B. (1983). Capacity demands in short-term memory for synthetic and natural word lists. *Human Factors, 83,* 17–32.

Massaro, D. W., & Light, J. (2004). Using visible speech for training perception and production of speech for hard of hearing individuals. *Journal of Speech, Language, and Hearing Research, 47,* 304–320.

Mayo, C., Scobbie, J. M., Hewlett, N., & Waters, D. (2003). The influence of phonemic awareness development on acoustic cue weighting strategies in children's speech perception. *Journal of Speech, Language, and Hearing Research, 46,* 1184–1196.

McCandliss, B., Fiez, J. A., Protopapas, A., Conway, M., & McClelland, J. L. (2002). Success and failure in teaching the [r]–[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective, & Behavioral Neuroscience, 2,* 89–108.

McGarr, N. S. (1983). The intelligibility of deaf speech to experienced and inexperienced listeners. *Journal of Speech and Hearing Research, 26,* 451–458.

McNaughton, D., Fallon, D., Tod, J., Weiner, F., & Neisworth, J. (1994). Effects of repeated listening experiences on the intelligibility of synthesized speech. *Augmentative and Alternative Communication, 10,* 161–168.

Mirenda, P., & Beukelman, D. R. (1987). A comparison of speech synthesis intelligibility with listeners from three age groups. *Augmentative and Alternative Communication, 5,* 84–88.

Mirenda, P., & Beukelman, D. R. (1990). A comparison of intelligibility among natural speech and seven speech synthesizers with listeners from three age groups. *Augmentative and Alternative Communication, 6,* 61–68.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology, 31,* 477–488.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57.

Nusbaum, H. C., & Pisoni, D. B. (1985). Constraints on the perception of synthetic speech generated by rule. *Behavior Research Methods, Instruments, & Computers, 17,* 235–242.

Nusbaum, H. C., & Schwab, E. C. (1986). The role of attention and active processing in speech perception. In E. C. Schwab & H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines: Vol. 1. Speech perception* (pp. 113–157). San Diego, CA: Academic Press.

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America, 97*, 593–608.

Pisoni, D. B., & Luce, P. A. (1987). Trading relations, acoustic cue integration, and context effects in speech perception. In M. E. H. Schouten (Ed.), *The psychophysics of speech perception* (pp. 155–172). Dordrecht, the Netherlands: Martinus Nijhoff Publishing.

Quinn, P. C., Schyns, P. G., & Goldstone, R. L. (2006). The interplay between perceptual organization and categorization in the representation of complex visual patterns by young infants. *Journal of Experimental Child Psychology, 95,* 117–127.

Rabbitt, P. (1991). Mild hearing loss can cause apparent memory failures which increase with age and reduce with IQ. *Acta Otolaryngologica (Stockholm), 476*(Suppl.), 167–176.

Repp, B. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin, 92,* 81–110.

Reynolds, M. E., Isaacs-Duvall, C., & Haddox, M. L. (2002). A comparison of learning curves in natural and synthesized speech comprehension. *Journal of Speech, Language, and Hearing Research, 45,* 802–810.

Reynolds, M. E., Isaacs-Duvall, C., Sheward, B., & Rotter, M. (2000). Examination of the effects of listening practice on synthesized speech comprehension. *Augmentative and Alternative Communication, 16,* 250–259.

Romski, M. A., & Sevcik, R. A. (1996). *Breaking the speech barrier: Language development through augmented means*. Baltimore: Brookes.

Rousenfell, S., Zucker, S. H., & Roberts, T. G. (1993). Effects of listener training on intelligibility of augmentative and alternative speech in the secondary classroom. *Education and Training in Mental Retardation, 12,* 296–308.

Samuel, A. G., & Newport, E. L. (1979). Adaptation of speech by nonspeech: Evidence for complex acoustic cue detectors. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 546–578.

Schepis, M. M., & Reid, D. H. (1995). Effects of voice output communication aid on interactions between support personnel and an individual with multiple disabilities. *Journal of Applied Behavior Analysis, 28,* 73–77.

Schmidt-Nielsen, A. (1995). Intelligibility and acceptability testing for speech technology. In A. K. Syrdal, R. W. Bennett, & S. L. Greenspan (Eds.), *Applied speech technology* (pp. 195–232). Boca Raton, FL: CRC Press.

Schneider, A., & Zuccolotto, A. (2002). *E-Prime user's guide.* Pittsburgh, PA: Psychology Software Tools.

Schwab, E. C., Nusbaum, H. C., & Pisoni, D. B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors, 27,* 395–408.

Schyns, P. G. (1998). Diagnostic recognition: Task constraints, object information, and their interaction. *Cognition, 67,* 147–179.

Schyns, P. G., Goldstone, R. L., & Thibaut, J.-P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences, 21,* 1–54.

Schyns, P. G., & Oliva, A. (1997). Flexible, diagnosticity-driven, rather than fixed, perceptually determined scale selection in scene and face recognition. *Perception, 26,* 1027–1038.

Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 682–696.

Shepard, R. N. (1972). Psychological representations of speech sounds. In E. E. David & P. B. Denes (Eds.), *Human communication: A unified view* (pp. 67–111). New York: McGraw-Hill.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review, 84,* 127–190.

Soli, S. D., & Arabie, P. (1979). Auditory vs. phonetic accounts of observed confusions between consonant phonemes. *The Journal of the Acoustical Society of America, 66,* 46–59.

Soli, S. D., Arabie, P., & Carroll, J. D. (1986). Discrete representation of perceptual structure underlying consonant confusions. *The Journal of the Acoustical Society of America, 79,* 826–837.

Spitzer, S. M., Liss, J. M., Caviness, J. N., & Adler, C. (2000). An exploration of familiarization effects in the perception of hypokinetic and ataxic dysarthric speech. *Journal of Medical Speech-Language Pathology, 8,* 285–293.

Statsoft. (2003). Statistica [computer software]. Tulsa, OK: Author.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America, 111,* 1872–1891.

Teoh, S. W., Neuburger, H. S., & Svirsky, M. A. (2003). Acoustic and electrical pattern analysis of consonant perceptual cues used by cochlear implant users. *Audiology and Neuro-Otology, 8,* 269–285.

Tjaden, K. K., & Liss, J. M. (1995). Listener familiarity in the perception of dysarthric speech. *Clinical Linguistics & Phonetics, 9,* 139–154.

Tjaden, K., Rivera, D., Wilding, G. E., & Turner, G. (2005). Characteristics of the lax vowel space in dysarthria. *Journal of Speech, Language, and Hearing Research, 48,* 554–566.

Tun, P. A., & Wingfield, A. (1994). Speech recall under heavy load conditions: Age, predictability, and limits on dual-task interference. *Aging and Cognition, 1,* 29–44.

Venkatagiri, H. S. (1994). Effects of sentence length and exposure on the intelligibility of synthesized speech. *Augmentative and Alternative Communication, 10,* 96–104.

Received July 5, 2006

Revision received November 6, 2006

Accepted April 4, 2007

DOI: 10.1044/1092-4388(2007/100)

Contact author: Alexander L. Francis, Speech, Language, and Hearing Sciences, Purdue University, Heavilon Hall, 500 Oval Drive, West Lafayette, IN 47907.
E-mail: francisa@purdue.edu.

Appendix A (p. 1 of 3). Confusion matrices for all three outcome groups (strong learners, weak learners, and untrained controls) on both the pretest and the posttest.

Table A-1. Pretest confusion matrix (strong learners).

Stimulus presented	Number of each type of response													
	b	d	f	g	k	m	n	p	s	t	v	w	y	z
b	12	189	0	85	1	0	0	0	0	5	0	0	0	0
d	0	216	0	21	0	0	2	0	2	3	0	0	6	34
f	1	0	136	1	0	3	1	8	146	1	1	0	0	0
g	28	94	0	114	0	0	2	0	0	0	0	0	19	0
k	0	5	0	12	67	0	0	66	1	15	1	0	8	0
m	0	1	5	1	0	13	3	3	0	1	112	17	13	11
n	0	1	6	1	0	30	17	0	1	0	148	9	17	19
p	4	50	0	33	57	0	0	23	0	49	0	0	0	0
s	0	0	0	0	0	0	0	0	299	0	0	0	0	2
t	1	30	0	12	3	0	0	11	22	79	0	0	0	121
v	1	1	33	1	1	1	2	0	10	1	83	1	2	113
w	1	0	0	1	0	28	1	0	0	0	9	230	11	0
y	0	0	0	0	0	1	0	0	0	0	0	0	288	0
z	0	0	0	1	0	0	1	0	5	0	0	0	14	271

Note. Different rows have different total numbers of responses because not all participants gave a response to every stimulus presentation, and some responses were uninterpretable or indicated the perception of sounds not among the set presented.

Table A-2. Posttest confusion matrix (strong learners).

Stimulus presented	Number of each type of response													
	b	d	f	g	k	m	n	p	s	t	v	w	y	z
b	80	163	0	53	7	0	0	0	0	2	0	0	0	0
d	0	292	0	9	0	0	0	0	0	2	0	0	0	4
f	0	0	250	0	0	0	0	2	48	0	1	1	0	0
g	11	79	0	170	0	0	1	0	1	0	0	1	6	0
k	0	1	0	2	67	0	0	57	0	3	0	0	8	0
m	1	0	4	0	0	69	13	0	0	0	91	20	6	0
n	0	0	1	0	0	77	70	0	0	1	113	0	3	0
p	5	15	0	22	31	0	0	67	1	69	0	0	0	1
s	0	0	1	0	0	0	0	0	309	0	0	0	0	0
t	0	32	1	5	0	0	0	6	20	180	0	0	0	45
v	0	0	9	2	0	2	4	0	3	0	156	1	0	100
w	0	1	0	1	0	32	4	0	0	0	6	249	0	0
y	0	0	0	0	0	1	0	0	0	0	1	0	299	0
z	1	1	0	1	0	1	0	0	11	0	0	0	7	278

Appendix A (p. 2 of 3). Confusion matrices for all three outcome groups (strong learners, weak learners, and untrained controls) on both the pretest and the posttest.

Table A-3. Pretest confusion (weak learners).

Stimulus presented	Number of each type of response													
	b	d	f	g	k	m	n	p	s	t	v	w	y	z
b	9	137	0	12	0	0	0	1	0	5	0	0	0	0
d	0	150	0	2	0	0	0	0	1	0	0	1	0	4
f	0	1	85	1	0	0	0	1	68	9	0	0	0	0
g	10	95	0	48	0	0	0	0	0	0	0	0	5	0
k	0	6	0	0	31	0	0	48	0	12	0	1	1	4
m	0	0	1	3	0	13	18	0	1	0	38	15	3	7
n	0	3	0	2	0	14	33	0	8	0	50	8	0	2
p	0	58	0	21	22	0	0	12	0	21	1	1	0	0
s	0	0	0	1	0	0	0	0	162	0	0	0	1	1
t	0	30	0	6	4	0	0	0	16	22	0	0	0	54
v	0	3	2	5	0	0	5	0	10	9	33	2	0	77
w	0	0	0	0	0	13	2	0	0	0	0	112	21	0
y	0	1	0	0	0	0	0	0	0	0	0	1	162	0
z	0	5	0	2	0	0	0	0	9	1	0	0	0	138

Table A-4. Posttest confusion matrix (weak learners).

Stimulus presented	Number of each type of response													
	b	d	f	g	k	m	n	p	s	t	v	w	y	z
b	15	144	1	8	0	0	0	0	0	0	0	0	0	0
d	2	156	1	6	0	0	0	0	0	0	0	0	0	2
f	0	1	98	2	0	0	0	1	60	2	1	0	1	0
g	11	91	0	55	0	0	0	0	0	0	1	0	0	0
k	0	9	0	0	25	0	0	43	0	15	0	0	0	1
m	0	1	0	0	0	19	23	0	0	1	42	3	1	0
n	0	0	0	1	1	31	41	0	4	0	39	10	1	4
p	1	32	1	3	22	0	0	19	0	52	0	0	1	1
s	1	3	0	0	0	0	0	0	162	0	0	0	0	0
t	1	25	0	5	4	0	0	0	9	45	0	0	1	47
v	0	0	6	2	0	0	0	0	7	10	36	1	0	78
w	0	0	0	0	0	29	2	0	0	1	1	107	12	0
y	0	0	0	0	1	0	0	0	0	0	0	0	165	0
z	0	1	0	3	0	0	0	0	7	0	0	0	1	145

Appendix A (p. 3 of 3). Confusion matrices for all three outcome groups (strong learners, weak learners, and untrained controls) on both the pretest and the posttest.

Table A-5. Pretest confusion matrix (untrained controls).

Stimulus presented	Number of each type of response													
	b	d	f	g	k	m	n	p	s	t	v	w	y	z
b	0	14	2	4	1	2	0	0	8	0	8	7	5	8
d	0	19	3	7	0	0	0	0	11	2	4	3	8	6
f	0	18	3	9	0	0	0	0	12	2	0	8	7	4
g	0	18	3	6	0	2	0	0	15	5	1	0	4	13
k	0	13	3	7	1	1	0	0	9	3	3	8	8	8
m	0	18	3	4	1	0	0	0	10	5	2	8	7	10
n	0	20	4	3	0	1	0	0	19	1	0	4	10	5
p	0	26	1	7	2	1	0	0	10	3	3	5	5	6
s	0	10	1	10	2	0	0	0	7	7	1	4	9	7
t	0	16	1	9	0	0	0	0	13	4	5	3	6	7
v	0	14	3	1	2	2	1	0	19	4	3	2	6	6
w	0	18	2	7	1	0	1	0	17	4	1	5	3	5
y	0	16	1	2	1	2	0	0	10	0	2	6	6	7
z	0	19	3	3	1	0	1	0	9	3	3	6	8	5

Table A-6. Posttest confusion matrix (untrained controls).

Stimulus presented	Number of each type of response													
	b	d	f	g	k	m	n	p	s	t	v	w	y	z
b	0	53	0	8	0	0	0	0	0	1	0	8	0	1
d	0	43	0	0	0	0	0	0	19	0	0	0	2	17
f	0	20	0	0	0	0	0	0	26	0	0	8	9	9
g	0	27	8	0	0	0	0	0	9	0	0	1	9	9
k	0	10	0	0	0	0	0	0	16	0	0	9	9	13
m	0	42	0	0	0	0	0	0	18	2	0	0	9	9
n	0	35	0	0	0	0	0	0	2	0	0	0	9	18
p	0	49	0	0	0	0	0	0	2	1	0	0	9	13
s	0	27	0	0	0	0	0	0	18	0	0	1	1	11
t	0	34	0	0	0	0	0	0	8	1	0	2	1	33
v	0	12	0	0	0	0	0	0	17	2	0	1	0	26
w	0	34	0	0	0	0	0	0	18	1	0	2	8	9
y	0	19	0	0	0	0	0	0	10	1	0	8	24	1
z	0	24	0	0	0	0	0	0	9	2	0	2	0	25