Journal of Experimental Psychology:
Human Perception and Performance
2002, Vol. 28, No. 2, 349–366

0096-1523/02/$5.00 DOI: 10.1037//0096-1523.28.2.349

Selective Attention and the Acquisition of New Phonetic Categories

Alexander L. Francis
University of Hong Kong

Howard C. Nusbaum
University of Chicago

A class of selective attention models often applied to speech perception is used to study effects of training on the perception of an unfamiliar phonetic contrast. Attention-to-dimension (A2D) models of perceptual learning assume that the dimensions that structure listeners' perceptual space are constant and that learning involves only the reweighting of existing dimensions to emphasize or de-emphasize different sensory dimensions. Multidimensional scaling is used to identify the acoustic–phonetic dimensions listeners use before and after training to recognize the 3 classes of Korean stop consonants. Results suggest that A2D models can account for some observed restructuring of listeners' perceptual space, but listeners also show evidence of directing attention to a previously unattended dimension of phonetic contrast.

Recently, speech researchers have begun to make use of perceptual classification models that stem from the generalized context model (GCM) of perceptual learning and categorization developed by Nosofsky (1986). This model has particular application to phonetic learning (acquisition of new phonetic categories) in the context of first and second language acquisition (e.g., see Jusczyk, 1994, 1997; Kuhl & Iverson, 1995; Pisoni, 1997), although it is usually applied as a post hoc explanation of experimental results. This model basically assumes that categorization can be understood within a spatial metaphor (see Shepard, 1957, 1974; but also Tversky, 1977; Tversky & Gatti, 1982) in which sensory attributes of stimuli are represented as the dimensional structure of a categorization space. In broad terms, learning shifts attention to dimensions relevant for classification and away from dimensions that are irrelevant. The operations of attending and ignoring are formalized as a stretching or shrinking of the dimensions to represent shifts of attention to or away from dimensions of categorization. The GCM framework seems to fit with some general patterns of findings in perceptual learning of speech (see Pisoni, Lively, & Logan, 1994). More importantly, the GCM formalizes a theory of selective attention, and therefore applying it to phonetic learning provides a concrete cognitive model to describe phenomena that are commonly termed *attentional* without further clarification (see especially discussions by Jusczyk, 1994; Pisoni et al., 1994).

Although most speech researchers who invoke cognitive models of selective attention typically cite Nosofsky (1986), some recent speech results (Iverson & Kuhl, 1995) are more suggestive of a different but related model of selective attention, exemplified by the theory developed by Goldstone (1993, 1994). Both the GCM model and Goldstone's model share many characteristics that make them desirable to speech researchers, and, based on their similarities, these two models could be collectively termed *attention-to-dimension* models, or A2D models. Shared characteristics include the assumption of a spatial metaphor and an emphasis on changes in the distribution of selective attention as the principal mechanism of perceptual learning. Of particular interest for our purposes, both Nosofsky and Goldstone characterize this mechanism in terms of adjusting the attentional weight given to individual dimensions of contrast. Although these models formally incorporate attention as the weighting mechanism, this basic concept of categorization through *dimensional warping* is shared by a large class of models, including Kuhl's prototype-based perceptual magnet (Iverson & Kuhl, 1995, 1996; Kuhl & Iverson, 1995) and various connectionist models based on neural map formation (e.g., Guenther, Husain, Cohen, & Shinn-Cunningham, 1999; Kruschke, 1992; McClelland, 2001).

In A2D warping models, learning is treated in terms of a pair of complementary attentional operations that serve to change the structure of perceptual space to produce categorization. These operations are formalized in terms of a weight or multiplier that stretches or shrinks the dimensions of perceptual contrast that structure perceptual space. Focusing attention on a particular sensory dimension increases the multiplier of that dimension, in effect stretching it, making the differences between any two (nonidentical) points along that dimension appear greater (because the distance, and thus the difference, between them has increased). Conversely, withdrawing attention from a dimension causes that dimension to shrink, because differences between points along that dimension are reduced. Although this is a small set of attentional

Alexander L. Francis, Department of Speech and Hearing Sciences, University of Hong Kong, Hong Kong SAR, China; Howard C. Nusbaum, Department of Psychology, University of Chicago.

Material in this article derives from part of a doctoral dissertation submitted by Alexander L. Francis to the Department of Psychology and the Department of Linguistics at the University of Chicago. This work was supported in part by a grant from the Division of the Social Sciences at the University of Chicago to Howard C. Nusbaum. We are grateful to Won-Seok Cho, Valter Ciocca, Elaine J. Francis, Rachel Hemphill, Anne Henly, Janellen Huttenlocher, Karen Landahl, David McNeill, Terry Regier, Steve Shevell, and three anonymous reviewers for their helpful comments and advice on earlier versions of this work.

Correspondence concerning this article should be addressed to Alexander L. Francis, Department of Speech and Hearing Sciences, 5/F Prince Philip Dental Hospital, 34 Hospital Road, Hong Kong SAR, China. E-mail: afrancis@hkusua.hku.hk

operations, thus far they have proved sufficient to account for many aspects of perceptual learning in the laboratory.

With these attentional operations, all dimensional warping models are capable of modeling fundamental aspects of category learning, including acquired distinctiveness between categories and acquired equivalence (similarity) within categories, as described by Gibson (1969; see also Goldstone, 1998). Specific A2D warping models differ, however, in the particular implementation of these operations. For example, according to the GCM, attention can only stretch or shrink a dimension uniformly over its entire span. Such a mechanism would be unable to accomplish concomitant stretching around category boundaries (acquired distinctiveness, reflecting increasing between-categories sensitivity) and shrinking around category prototypes (acquired similarity, reflecting decreasing within-category sensitivity) along a single dimension of contrast. The same would be true of connectionist models in which dimensional weights are modeled as connection strengths (multipliers) in a simple feedforward network. However, results described by Kuhl and Iverson (1995, summarizing results presented by Iverson & Kuhl, 1995) suggest that such combinations of stretching and shrinking along a single dimension are characteristic of phonetic learning, although Iverson and Kuhl (2000) argued that category boundary effects (stretching) and prototype effects (shrinking) arise from the operation of distinct mechanisms. Iverson and Kuhl (1995) found that tokens consistently identified as good exemplars of the categories /i/ and /e/ cluster together (around their respective category prototypes) in perceptual space. In contrast, intermediate tokens lying between these two clusters of good tokens appear to be much farther apart in perceptual space, although all tokens were equally separated in acoustic space. In other words, tokens that are acoustically similar to category prototypes are moved closer to the prototype through adjustments of the perceptual space, whereas tokens that are far from category prototypes are perceived as being even more different. Similar observations of localized stretching and shrinking have been described in other domains of perceptual categorization (Goldstone, 1993, 1994; but see Livingston, Andrews, & Harnad, 1998), giving rise to a kind of model that, while still fundamentally a dimensional warping model, might be more accurately described as *localized warping* because it is specifically designed to accommodate differential warping along the same dimension of contrast (see also Guenther et al., 1999, for a connectionist model, which, while in many ways different from Goldstone's, is in this respect fundamentally a localized warping model).[1]

Iverson and Kuhl's (1995) results suggest that localized warping may be the preferable dimensional warping model to account for category learning effects in speech perception. However, it is not clear that the current specifications of dimensional warping models are sufficient to account for all the details of other recent studies in speech perception. Dimensional warping models of perceptual learning were developed primarily within the context of studies using simple visual or auditory stimuli specifically created for the experiment (e.g., Goldstone, 1994; Guenther et al., 1999; Nosofsky, 1986). Thus far, perceptual learning studies have typically used artificial and arbitrary categories and extremely simple stimuli. In these studies, the formation of a category is essentially a matter of picking and choosing between the dimensions of contrast that the experimenter has selected. The only dimensions available for categorization are those that the experimenter has chosen for investigation and therefore built into the stimuli, and there is no necessary assumption that listeners have any category-level system for organizing those dimensions before the experiment begins.

In contrast, the speech signal is richer in information and typically provides multiple, mutually reinforcing (integral), but also potentially redundant (and recombinant) cues to phonological contrasts (e.g., Nittrouer & Miller, 1997; Repp, 1982). Furthermore, in adult phonological acquisition, listeners come to the task equipped with a complex, ecologically valid knowledge system for categorizing speech sounds. Listeners' native language system strongly influences their subsequent perception of speech such that, for example, some unfamiliar phonological contrasts are quite easy to learn, whereas others are extremely difficult (Best, McRoberts, & Sithole, 1988; Burnham, 1986; Polka, 1991, 1992; Strange, 1995; Werker & Tees, 1984). In other words, from the perspective of an A2D warping model of perceptual learning, adult listeners already possess a structured perceptual space mapping auditory stimuli onto categorical knowledge, and this structure can be expected to influence learning in predictable ways.

Best and her colleagues (Best, 1994, 1995; Best et al., 1988) have developed a taxonomy of four types of cross-language contrasts that builds on this observation, and two of these types are of particular interest here. In the case of *single category* (SC) contrasts, two (or more) foreign categories map equally well to a single native category, although both may be heard as strange or discrepant versions of the single native category. In the case of contrasts that depend on *category goodness* (CG), two foreign categories map to a single native category, but they do so to differing degrees.

Within a dimensional warping model, we can investigate the different predictions these two contrasts make for learning. Specific foreign categories in a CG contrast differ acoustically in a way that causes them to map unequally onto a single native category in a listener's existing perceptual space. The acoustic properties that distinguish them to the nonnative listener (regardless of whether these acoustic properties are the same as those used by native speakers of the foreign language) may be allophonic in the native language or they may be highly correlated with (i.e., integral with) properties that are not distinctive with respect to the native categories. In either of these two cases, listeners have some experience with the properties that must be used to distinguish the two foreign categories, so it should be possible for listeners to learn to distinguish CG contrasts, either by increasing attention to the underattended dimension of allophonic distinction or by separating a previously integral set of correlated dimensions. In contrast, it would appear that the only way to learn an SC contrast would be to learn to attend to a new dimension of contrast, because no currently attended dimension provides sufficient information to qualitatively distinguish the two foreign categories; the dimensions that distinguish an SC contrast are irrelevant to native contrasts

[1] It should be noted that Goldstone (1994) did not observe acquired equivalence along any categorization-relevant dimension, although there was one case of acquired equivalence along a categorization-irrelevant dimension. However, the degree of acquired distinctiveness along categorization-relevant dimensions was smaller within categories than between. This could be taken as evidence of the interaction of (weaker) within-category (local) acquired equivalence with (stronger) global sensitization of the entire dimension.

and are thus ignored by native phonetics. In this case, listeners have to locate and attend to a dimension that was previously unattended because of the developmentally acquired constraints of the native phonology. In other words, although phonetic learning may involve shifting attentional weight between existing dimensions of contrast (e.g., as suggested by Francis, Baldwin, & Nusbaum, 2000; Nittrouer & Miller, 1997), it may also involve the induction of a completely new dimension to acquire an SC contrast, as well as the integration or separation of existing dimensions, in either case forming new dimensions that are more functional in the foreign phonetic system. This would be akin to the developmental proposal made by Smith and Kemler (1977) in which integral dimensions may be formed by attention through perceptual learning.

Attention has often been invoked to account for phonological acquisition, and dimensional warping models are often suggested as post hoc possibilities to account for the effects of phonetic learning (e.g., Iverson & Kuhl, 1995; Jusczyk, 1994, 1997; Nusbaum & Goodman, 1994; Nusbaum & Lee, 1992; Pisoni et al., 1994). However, although it is commonly accepted that learning new phonological contrasts may involve learning to attend to a new phonetic dimension, studies of adult phonological learning have tended to minimize the possibility that participants might learn to attend to new dimensions of phonetic contrast. Two of the more commonly studied cases of adult phonological learning involve the acquisition of contrasts that are not, strictly speaking, novel to learners. For example, in the synthesized Thai stimuli used by Pisoni and his colleagues, voice onset time (VOT) is the only distinguishing acoustic cue (McClaskey, Pisoni, & Carrell, 1983; Pisoni, Aslin, Perey, & Hennessy, 1982). This contrast is clearly a CG contrast, as prevoiced stimuli are perceptibly different from unvoiced stimuli, even for naïve English listeners, as demonstrated in the discrimination data prior to training reported by Pisoni et al. (1982). Furthermore, for English speakers, learning to separate [b] from [p] (which is already distinguishable from [p^h] according to VOT) merely requires that listeners learn to make a new category distinction along an already attended dimension of contrast (VOT).[2]

Similarly, the acquisition of the English /r/–/l/ distinction by native speakers of Japanese (Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999; Iverson & Kuhl, 1996; Lively, Pisoni, & Logan, 1992; Yamada, 1995; Yamada & Tohkura, 1992), while more likely to be an SC contrast, can also apparently be learned without recourse to attending to a new dimension of phonetic contrast. Indeed, it probably requires that listeners learn to ignore a previously attended dimension. Whereas English-speaking listeners in Yamada and Tohkura's (1992) experiments distinguished /r/ from /l/ almost exclusively on the basis of differences in the center frequency of the third formant (F3; low for /r/, higher for /l/), Japanese listeners made their category decisions on the basis of a combination of F3 and the second formant frequency (F2) cues. Thus, for Japanese listeners, learning to distinguish /r/ from /l/ involves not only learning to pay more attention to the (already somewhat attended) F3 cue but also to ignore unhelpful information about F2.

To investigate the acquisition of a new dimension of phonetic contrast, one must use a contrast made along an acoustic dimension that is not linguistically distinctive in the listeners' native language; that is, either an SC contrast that requires learning to attend to a completely unfamiliar dimension or a CG contrast that involves separating an integral dimension. Completely unfamiliar SC contrasts are quite difficult to find, because even cross-linguistically rare contrasts such as the Hindi dental–retroflex stop contrast may correspond to allophonic distinctions in another language. For example, although both the Hindi dental [t] and retroflex [t] assimilate very clearly to the single native English category /t/ (Werker & Logan, 1985), English does contrast dental with alveolar place of articulation in fricatives (e.g., in the words *thin* vs. *sin*), and retroflex (and possibly dental) stops can appear allophonically as a consequence of coarticulation, for example, retroflex before /r/, as in *trip* and *drip* (Polka, 1991). Thus, although the contrast does not itself appear in English, some of the acoustic cues that signal this contrast in Hindi may in fact be familiar to English listeners. Despite this, it has proved extremely difficult to train English listeners to hear a dental–retroflex stop contrast in the laboratory (Polka, 1991; Tees & Werker, 1984), possibly indicating that English listeners are not used to attending to the acoustic cues that signal this contrast in Hindi. However, the reported difficulty of training this contrast makes it less than ideal for the purposes of the present article. An example of the second sort of contrast would be one that is comparatively easily learned by English speakers (unlike the Hindi retroflex–dental stop contrast) but is still not made along an acoustic dimension that is known to be of primary linguistic importance in English. Such a dimension should be one that covaries with other, more salient cues and is therefore treated as integral with those other cues. The three-way voicing distinction found in Korean syllable-initial stop consonants fits this characterization. Unlike the VOT-based stop contrast found in Thai, stop consonants in Korean are generally described as differing along at least two distinctive dimensions (e.g., Kang, 1998; Schmidt, 1996) for native speakers. The exact feature specification of these three consonant classes is often debated, and it is not within the scope of this article to do more than note the existence of this issue.[3] We adopt the terminology and transcription used by Han and Weitzman (1970). Thus, the three kinds of stops in this study are the following: aspirated, /p^h/, /t^h/, and /k^h/; weak, /p/, /t/, /k/; and strong, /P/, /T/, and /K/. Collectively, these categories are often considered to differ according to voicing features,[4] and this terminology is relatively uncontroversial. The three classes of stops do not contrast in all positions within the syllable in Korean, but they are realized distinctively in initial position. For example, Han and Weitzman (1965) listed the words [p^hul] *grass* versus [Pul] *horn* versus [pul] *fire*; [t^hal] *mask* or *trouble, problem* versus [Tal] *daughter* versus [tal] *moon*; and

[2] Note that we are not aware of any study that demonstrates that English-speaking listeners necessarily attend to VOT cues when making a voicing distinction in natural speech. However, there is considerable evidence that such cues are clearly usable when present in stimuli in which all other cues have been neutralized (Lisker & Abramson, 1970).

[3] In fact, most of the phonological debate involves how to deal with the neutralization of (aspects of) this contrast in medial and final positions. In syllable initial position, the tripartite nature of the contrast is not in debate.

[4] Note that Hardcastle (1973) considers the aspirated stops to be strong as well, on the basis of their patterning with strong consonants in the acoustic parameters we refer to here as RISE and f_0 onset. The issue of phonological specification is not of primary concern in this article and can safely be ignored.

[k^hida] *keep pets* or *to play a stringed instrument* versus [Kida] *insert* versus [kida] *crawl.*

Acoustically, the distinction is not as easily defined. Most researchers find some overlap in VOT between categories, particularly between the weak and strong stops (Han & Weitzman, 1970; Lisker & Abramson, 1964), although Hardcastle (1973) found no such overlap between any categories. A number of other acoustic features have been described as differing systematically between weak and strong stops in Korean, including the rate of increase in vowel amplitude (which we call RISE), such that aspirated and weak consonants have a longer RISE than do strong consonants (Han & Weitzman, 1970; Hardcastle, 1973; Lisker & Abramson, 1964). Similarly, both the fundamental frequency (f_0) and the clarity of formant structure at the onset of phonation (CLEAR) have been related to the same distinction, such that vowels following weak consonants have a more damped quality (lower values of CLEAR) and a lower onset f_0 (Han & Weitzman, 1970; Hardcastle, 1973).

Based on previous studies of the perception of English consonants, we know that native speakers of English attend to VOT in making decisions about stops. Less clear is whether they will attend to onset f_0 or not. Onset f_0 does covary with other cues to voicing in English stop consonant production, and it has been demonstrated that onset f_0 can function as a sufficient cue to the perception of voicing contrasts in the absence of other cues, at least for some listeners (Haggard, Ambler, & Callow, 1970). This suggests that American listeners may be aware that f_0 can play a role in the voicing specification of stop consonants, but they do not easily treat it as distinct from other features that cue voicing. Under the assumption that American English listeners are most likely to be attending to VOT, it may be predicted that they will initially be able to distinguish the aspirated consonants from the other two categories using their phonetic knowledge of voicing. Furthermore, if they do not attend to f_0 or CLEAR as dimensions separate from VOT on the pretest, then we may predict that they will not be able to distinguish between the weak and strong consonants. In this case, listeners unused to attending separately to f_0 or CLEAR will have to induce a new phonetic dimension by shifting their attention to this acoustic property to learn the Korean phonetic structure.

Our predictions further depend on the assumption that the stimuli used in this experiment exhibit patterns of acoustic features similar to those described by previous researchers, which, given the wide range of variation between previous results, need not be assumed. Experiment 1, while not intended as an exhaustive study of the acoustic features of Korean stop consonants, is designed to identify those acoustic features in our stimuli that are most likely to function as cues to the three-way voicing contrast in our stimuli. The results of Experiment 2 illustrate native Korean speakers' attentional distribution when listening to these same stimuli and provide a sense of the phonetic structures that trained nonnative speakers might be expected to learn. Finally, Experiment 3 is designed to investigate the changes that occur in nonnative listeners' mental representations of bilabial stop consonants as a consequence of learning to recognize three classes of consonants from Korean. The primary method of analysis in Experiments 2 and 3 is multidimensional scaling (MDS), which is used to develop a spatial representation of the listener's phonetic space before and after training.

In Experiment 2, MDS is used to identify the phonetic dimensions that native Korean speakers attend to when distinguishing three classes of Korean stop consonants. In Experiment 3, the same techniques are applied to investigate the phonetic dimensions attended to by native speakers of American English before and after they are trained to recognize the same three classes of consonants. Separate MDS solutions are calculated for the native speakers and for the trained participants' pretest and posttest to allow for the possibility that the optimum number of dimensions may differ as a consequence of linguistic experience (see Livingston et al., 1998). Within the framework of current A2D models of perceptual learning (including both the GCM and Goldstone's localized warping model), MDS can provide evidence relevant to investigating the attentional operations used by listeners during phonetic learning. By more closely examining these attentional operations, we can better understand how current A2D models can be used to explain phonetic learning. Furthermore, we are interested in documenting the redirection of attention to a dimension of phonetic contrast that does not appear to be attended to prior to training (e.g., f_0 or CLEAR), if in fact our English-speaking listeners show no evidence of attending to this dimension on the pretest. Such redirection of attention would constitute evidence for a phenomenon that is assumed to underlie certain kinds of phonetic learning but that has not been identified experimentally.

Experiment 1

As noted earlier, Korean initial stop consonants are described as differing across three categories of voicing: aspirated, weak, and strong. These three categories are described as being formed from two different acoustic dimensions, termed RISE and f_0–CLEAR. In the first experiment, we carried out an acoustic analysis of a set of naturally produced Korean initial stop consonants that would serve as the experimental stimuli in subsequent experiments. The purpose of the analysis is to determine the degree to which these stimuli conform to the previous reports of acoustic cue patterns distinguishing voicing among Korean initial stop consonants (e.g., Han & Weitzman, 1970; Hardcastle, 1973; Lisker & Abramson, 1964).

Method

Stimuli for this experiment consisted of five sets of syllables recorded by a male native speaker of Korean (Seoul dialect) who is experienced at teaching Korean as a foreign language. He was paid $30 for approximately 2 hr of recording and preparation time. For recording, the talker was seated in a sound-isolating booth and spoke into a microphone approximately 8 in. (20.3 cm) in front of his lips. Recording was accomplished with a Tascam DA-20 mk2 DAT recorder located outside the booth. Syllables were digitized on a SPARC workstation using the ESPS/Waves+ interface (Entropic Research Laboratory, Washington, DC). Stimuli were low-pass filtered at 5 kHz and digitized at a sampling rate of 11025 Hz with 16-bit quantization.

Stimuli consisted of a total of 27 consonant–vowel (CV) syllables. These were created by combining the three places of stop articulation (bilabial, dental, and velar) with the three voicing classes (aspirated, weak, and strong). These nine consonants were combined with three monophthongal vowels /a/, /i/, and /o/ (approximately as in the American English words *hop*, *heap*, and the first part of the diphthong in *hope*) to create a total of 27 syllables.

During the recording session, this list of 27 syllables was then shown to the talker through a window in the sound booth written on individual file cards. Each card had written on it 1 syllable in Hangul, the Korean script. Cards were displayed at a regular rate, and the talker was instructed to read each syllable as it was shown. The list of 27 syllables was spoken five times, in different orders of presentation. The talker was instructed to read two of the lists (Lists 2 and 3) very clearly, "as if to an American student learning Korean." The other three lists (Lists 1, 4, and 5) were spoken in a regular, conversational manner. Each syllable was produced as a single utterance.

Only results of analyses of the bilabial consonants are reported here, because these are the stimuli that we used in the two subsequent listening experiments. All stimuli were analyzed acoustically using GW Instruments' SoundScope II speech analysis package (GW Instruments, Inc., Somerville, MA). Four acoustic parameters were measured: VOT, RISE, f_0, and CLEAR.

VOT refers to voice onset time, in milliseconds, measured from the end of the burst release to the start of voicing (identified as the initial zero-crossing of the first period of the vowel, measured from the waveform), which is commonly related to voicing distinctions (Han & Weitzman, 1970; Hardcastle, 1973; Lisker & Abramson, 1964). f_0 refers to the measured fundamental frequency (measured using autocorrelation [Rabiner & Schafer, 1978] with a frame advance of 2 ms) at the onset of the vowel, which has been shown to correlate with the strong–weak distinction in Korean (Han & Weitzman, 1970; Hardcastle, 1973). RISE refers to the measured duration, in milliseconds, from onset of vowel formants (identified as the first voicing pulse identified on a wide band [450 Hz window of analysis] spectrogram) to the peak vowel amplitude (measured from the acoustic waveform), which is an attempt to quantify the impressionistic observation (Han & Weitzman, 1970) that vowels following strong stops rise more abruptly in intensity. Diffusion refers to the average difference in amplitude, in decibels, between the first two peaks of a linear predictive coding plot (14 coefficients, taken at the onset of the vowel, identified as the first identifiable period of the waveform) and the trough between them—an attempt to quantify Han and Weitzman's (1970) impressionistic observation that the formant patterns in wide-band spectrograms of vowels following weak consonants appear weakened.

Results and Discussion

Table 1 shows the values of the acoustic parameters described above measured for those syllables containing bilabial consonants used in testing. VOT, RISE, f_0, and CLEAR distinguish the three different voicing qualities relatively well. VOT is quite good at distinguishing all three classes, such that strong consonants have the shortest VOT, followed by weak consonants with intermediate VOT, and aspirated consonants with quite long VOTs. The pattern for CLEAR is also obvious. Aspirated stops have the highest values of CLEAR, followed by strong stops, and finally weak stops. Although this pattern is consistent with the observations of Han and Weitzman (1970), it should be noted that CLEAR is likely to vary significantly as a consequence of background noise and may not therefore be a good candidate for a general (context-independent) phonetic feature. f_0 is also relatively good at distinguishing all three classes, with low values of f_0 corresponding to weak consonants, higher values for strong consonants, and marginally higher values for aspirated consonants. Finally, the picture for RISE is least obvious. RISE appears to be best for distinguishing the strong consonants (low RISE) from the weak and aspirated consonants (higher RISE).

Table 1
Acoustic Parameters Measured for Test Stimuli

Consonant	VOT (ms)	RISE (ms)	f_0 (Hz)	CLEAR (dB)
$[p]_1$	50	40	84	4.5
$[p]_4$	23	39	88	4.0
$[p^h]_1$	114	40	105	5.5
$[p^h]_4$	75	25	106	5.5
$[P]_1$	6	9	103	5.0
$[P]_4$	11	13	103	5.0

Note. VOT = voice onset time; RISE = measured duration from onset of vowel formants to peak vowel amplitudes; f_0 = measured fundamental frequency; CLEAR = clarity of formant structure at onset of phonation.

On the basis of this overall analysis, we might expect that the most useful acoustic features for distinguishing between these 18 stop consonants will be VOT, CLEAR, and possibly f_0. Within a spatial metaphor of categorical perception, we consider a cue to be sufficient for distinguishing between categories if the members of those categories can be linearly separated along that dimension (alone). As shown in Figure 1, the bilabial test tokens can be linearly separated according to VOT alone (and also according to f_0 and CLEAR, though the range of possible boundary values is much more tightly constrained). Furthermore, RISE is also a sufficient cue for distinguishing the strong from the aspirated and weak consonants, and thus in combination with f_0 or CLEAR could be used to distinguish between all three classes of stops.

Four acoustic parameters, identified on the basis of existing literature on the acoustic cues to the Korean stop consonant classes, appear to be good candidates for discriminating between the stimuli used here. Having identified these acoustic parameters, our next question is to determine how native Korean speakers use these parameters in making phonetic decisions. The fact that these parameters acoustically differentiate the phonological categories of Korean stops does not indicate whether these are the cues that Korean listeners attend to. Experiment 2 was carried out to examine the distribution of attention used by Korean listeners in classifying these stop consonants.

Experiment 2

In studying the perceptual learning of new phonetic contrasts, we must understand both native speakers' perceptual performance and the way that nonnative speakers' perceptions change. The acoustic analyses carried out in Experiment 1 provide an indication of the cues that listeners could possibly attend to in making Korean voicing decisions. However, the presence of cues does not guarantee that listeners actually make use of them (see Pickett, 1980). To understand how perceptual learning changes the phonetic space used by nonnative speakers during perception, we must understand how the phonetic space of native speakers is structured with respect to these cues. Our second experiment was designed to investigate how native Korean speakers make use of these cues in classifying the voicing of initial stop consonants. We used an MDS analysis to relate the structure of native Korean listeners' phonetic space to the acoustic cues described in Experiment 1.

Method

Participants. Five native speakers of Korean (3 male and 2 female) participated in this experiment. Three participants had lived in the United

States for less than 6 months at the time of the experiment and spoke only Korean at home. One participant had lived in the United States for slightly over a year and also spoke primarily Korean at home. The 5th participant was born in the United States but lived in Korea for 2 years as a child (age 4–5) and grew up speaking only Korean at home. However, at the time of the experiment, she used English as her primary language.

Stimuli. Stimuli used in this experiment were identical to those recorded and digitized for Experiment 1. Listeners were tested using syllables starting with bilabial, alveolar, and velar consonants, although only results involving bilabial consonants are analyzed here. For the difference rating task, participants heard half of all of the possible pairwise combinations of all syllables containing /a/ from Lists 1 and 4. The half that participants heard consisted of only those pairs beginning with syllables from List 1. For example, participants heard pairs $[p^ha]_1$–$[t^ha]_1$ and $[p^ha]_1$–$[k^ha]_4$ but not $[p^ha]_4$–$[t^ha]_1$ or $[p^ha]_4$–$[k^ha]_1$. Thus, there were a total of 162 pairs (two lists of three places of articulation and three classes of consonants is 18 possible syllables; every pairwise combination of these is 324 pairs, and half of that is 162 different pairs). For this experiment, only responses to stimuli containing bilabial consonants were analyzed. For the identification task, participants heard all syllables containing the vowel /a/ from Lists 1 and 4, for a total of 18 syllables (two lists of three places of articulation and three classes of consonants). Again, for this experiment only responses to syllables containing bilabial consonants were analyzed.

All stimuli were presented to participants binaurally at a comfortable listening level (approximately 70 dB peak sound pressure level [SPL]) over Sennheiser HD430 headphones in a sound-attenuated booth. Headphone level was under the control of each participant, but none chose to change it. Presentation of stimuli and collection of responses were digitally controlled on a SPARC workstation using a software interface.

Procedure. Participants attended two experimental sessions separated by at least 2 hr (and in four cases conducted on consecutive days). In the first session, the experimental procedure was explained to the participants. Also in this session, participants completed the first of two difference rating sessions. In the second session, they completed the second difference rating session and the identification task. Participants were paid $30 on completing the second session.

In each of the first and second difference rating session, participants were tested on two presentations of each pair of stimuli, for a total of four ratings per pair. The identification task consisted of 10 identification trials for each of the 18 syllables. All tokens in each task were presented in random order. Pairs of syllables in the difference rating task were separated by 250 ms of silence. Responses on each task were made as in Experiment 1. The only difference was that in the present experiment participants had a choice of nine possible pseudo-phonetic transcriptions. Each difference rating session was preceded by familiarization with one repetition of each syllable used in the pairs of stimuli. The identification task was also preceded by familiarization, that is, presenting each syllable twice while indicating the appropriate symbol for identification. Participants were also given a sheet of paper illustrating the transcription symbols and the corresponding characters in the Hangul script. Despite this, some participants reported having made a few errors owing to inexperience with the transcription system. Thus, identification scores may slightly underestimate perceptual performance. However, identification scores were almost perfect despite these few errors, averaging about 98% correct.

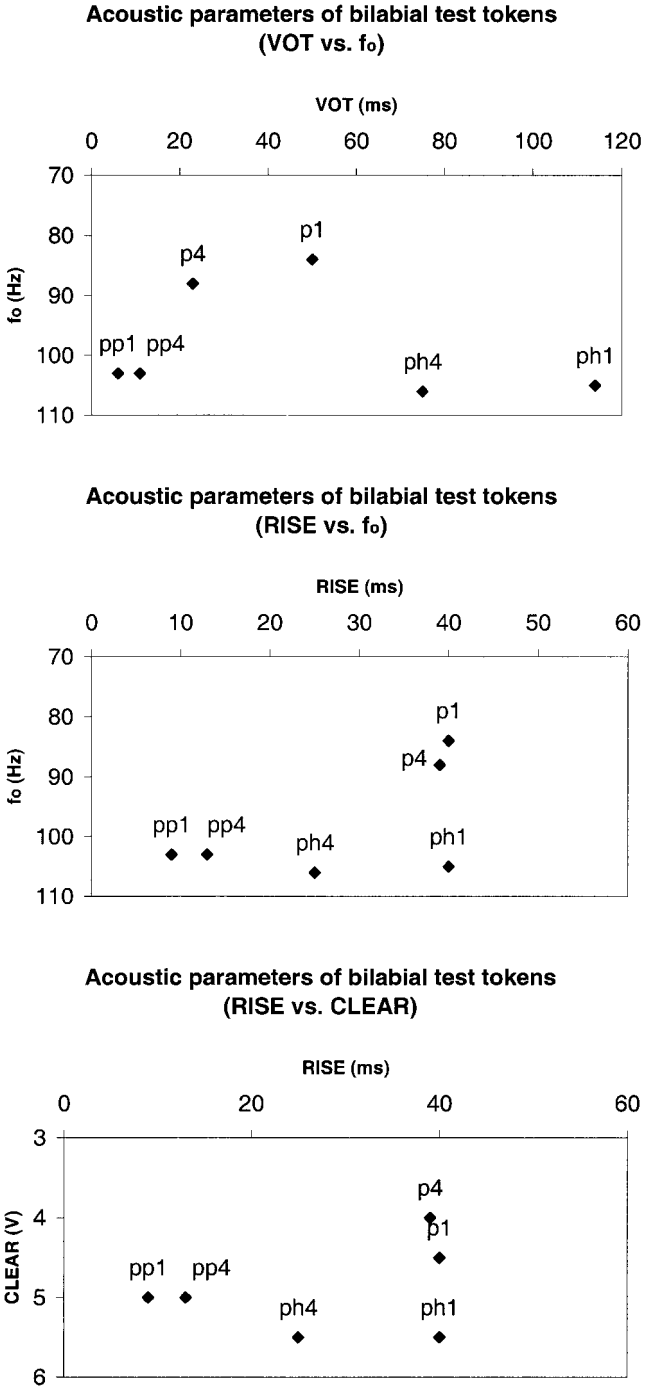

Figure 1. Plot of selected acoustic parameters (VOT, RISE, CLEAR [in volts], and f_0 at vowel onset) of bilabial test tokens (Experiment 1). Top: VOT is plotted along the horizontal axis, whereas f_0 is plotted inversely (increasing from top to bottom) along the vertical axis. Middle: RISE is plotted against f_0. Bottom: RISE is plotted against CLEAR. Two-dimensional plots were chosen to make more obvious the manner in which linear separability of voicing classes is facilitated in two dimensions, although it is possible for the single dimensions of VOT, f_0, and CLEAR. The inversion of the f_0 and CLEAR axes was chosen to facilitate comparison of this graph with subsequent graphs of the multidimensional scaling solutions generated from listeners' difference judgments involving these stimuli. VOT = voice onset time; f_0 = measured fundamental frequency; RISE = measured duration from onset of vowel formants to peak vowel amplitude; CLEAR = clarity of formant structure at onset of phonation.

Results and Discussion

Korean participants were extremely good at identifying the categories to which the stimuli belonged. The average percentage correct identification was 98% across all 5 participants, with a standard error of 1. Using participants' ratings of the degree of difference between pairs of consonants, we calculated an MDS solution using a three-way (individual-differences scaling) analysis.[5] Difference ratings were used because they are one of the most typical methods for estimating the perceptual similarity of stimuli for MDS analyses. Furthermore, Fox (1985) argued that, because paired-comparison judgments require listeners to remember stimuli before making a decision, paired-comparison judgments of speech signals require listeners to use both auditory (signal) information and linguistic (category) knowledge in a manner similar to that of normal speech perception. Thus, although making overt judgments about the similarity or difference of two speech sounds seems quite different from the process of normal speech perception, both tasks appear to draw on the same cognitive processes of memory and attention. Three-way MDS was used because the resulting axes are fixed by the input data (they are not subject to rotation) and are more likely to be interpretable or identifiable than those derived by two-way MDS (Kruskal & Wish, 1978).

Figure 2 shows the goodness of fit for solutions of varying dimensionality for native-speaker difference ratings on the bilabial consonants. In this case, there is a relatively clear elbow in the goodness-of-fit curve at two dimensions, and therefore a two-dimensional solution was initially calculated using the individual-differences scaling method implemented with the SAS MDS Procedure (SAS Institute, Inc., 1997). The resulting two-dimensional plot is shown in Figure 3.

As shown in Figure 3, the spatial distribution of tokens in the native listeners' solution space is similar to the distributions of tokens in acoustic space shown in Figure 1. From this figure, it appears that native speakers of Korean are indeed attending to those acoustic dimensions predicted by previous research and identified in the present stimuli. This impression is supported by the high degree of correlation between the location of tokens along the derived dimensions of the perceptual space and the location of tokens in the measured acoustic space, as shown in Table 2. Thus, the results of Experiments 1 and 2 suggest that when native speakers make phonetic decisions about the stimuli in these experiments, they are directing attention to both the VOT–RISE dimension of acoustic contrast and the f_0–CLEAR dimension.

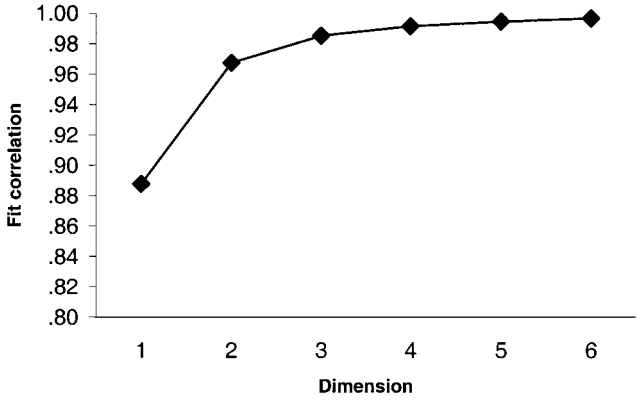

Figure 2. Fit correlation by dimensionality for native listeners' difference ratings on bilabial consonants only (Experiment 2).

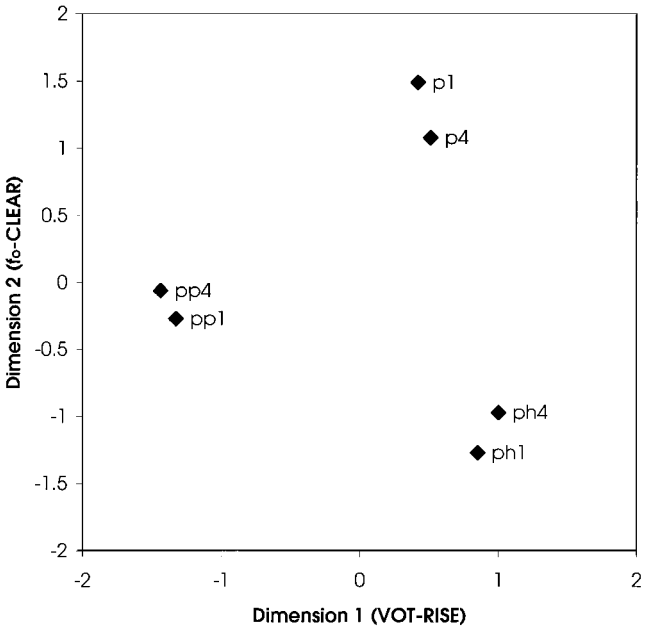

Figure 3. Native listeners' two-dimensional solution for bilabial stops. Tokens are transcribed as in Han and Weitzman (1970), with the exception that /P/ is written here as /pp/ and /p^h/ as /ph/. Numerals refer to the recitation list from which the token is drawn (see *Method* section of Experiment 1). CLEAR = clarity of formant structure at onset of phonation; f_0 = measured fundamental frequency; VOT = voice onset time; RISE = measured duration from onset of vowel formants to peak vowel amplitude.

Experiment 3

The third experiment was designed to investigate changes in nonnative listeners' mental representations of Korean stop conso-

[5] It must be noted that Korean participants heard only the top rectangular half of the matrix. Thus, there are no measured data points for pairs beginning with half of the syllables in the identification set. However, the MDS procedure is relatively robust and is designed to deal with situations in which one triangular half matrix of data is missing. In the ideal case in which there are measured values in both the upper triangular half matrix and the lower triangular half matrix (e.g., for both $[p]_1$–$[p^h]_4$ and $[p^h]_4$–$[p]_1$), the MDS procedure uses an average of the two. In cases in which one of the two triangular half matrices (or a particular cell from one triangular half matrix) is missing, the assumption of reflexivity—distance x–y is equivalent to distance y–x—provides a method for substituting existing values for missing ones. That is, because the similarity of $[p]_1$ to $[p^h]_4$ is assumed to be the same as the similarity of $[p^h]_4$ to $[p]_1$, the difference rating actually measured for pair $[p]_1$–$[p^h]_4$ can be substituted for the missing value of the pair $[p^h]_4$–$[p]_1$. It is only when there is a complete absence of values for either order of presentation that no approximation is possible. However, as long as the number of such completely missing values is relatively small (and in this case there are only six such completely missing values, of which only three are not pairs of identical tokens), doing without them merely adds slightly to the overall stress of the resulting solution.

Table 2
Correlations and p Values of Measured Acoustic Parameter Values With Locations of Tokens in Native Listeners' Perceptual Space (Bilabial Consonants Only)

	Dimension 1		Dimension 2	
Parameter	*r*	*p*	*r*	*p*
VOT	*.78*	*.065*	−.52	.287
RISE	**.84**	**.04**	.30	.562
f_0	−.21	.692	**−.95**	**.004**
CLEAR	.06	.905	**−.92**	**.010**

Note. Correlations significant at or below the $p = .05$ level are marked in bold. Nearly significant correlations ($p < .10$) are in italics. Stimulus values for all parameters for all tokens are shown in Table 1. VOT = voice onset time; RISE = measured duration from onset of vowel formants to peak vowel amplitudes; f_0 = measured fundamental frequency; CLEAR = clarity of formant structure at onset of phonation.

nants before and after training. On the basis of the results of previous research (e.g., Goldstone, 1994; Kuhl & Iverson, 1995; Livingston et al., 1998), we expect that same-category tokens will be perceived as more similar after training, whereas tokens from different categories will be perceived as more different after training. These results should be reflected in MDS analyses as a compression along particular dimensions within categories or expansion between categories. Furthermore, training is expected to induce listeners to attend to information not used prior to training. To correctly classify the stops in terms of Korean phonology, listeners will have to learn to make use of CLEAR or f_0, changing the dimensional structure of their perceptual space.

Thus, the main question in this study is whether or to what degree American English-speaking listeners attend to these acoustic cues when listening to these stimuli, and whether (or how) identification training will affect the distribution of nonnative speakers' attention. As VOT is typically considered the most salient cue to the English voiced–voiceless distinction, it is possible that American English-speaking listeners will attend primarily, or even exclusively, to this cue. The question of whether American English-speaking listeners will also attend to f_0 before training is an empirical one, but previous research suggests that they might. f_0 obviously plays a significant intonational role in English and can also function as a cue to the identification of stop consonants (Haggard et al., 1970), so in some contexts American listeners seem to attend to f_0, although not as a separate phonetic cue and probably not in the same way Korean listeners attend to it. Similarly, CLEAR may serve to distinguish breathy-voiced vowels and /h/ (as in *ahead*) from nonbreathy vowels.[6] However, Haggard et al. noted a great deal of between-listeners variation in the degree to which f_0 differences are sufficient to cue the perception of voicing differences. Because f_0 and VOT cues tend to pattern together in English, it is possible that listeners have learned to treat these two cues as integral components of a composite voicing cue that is only separable for some listeners, or with some difficulty. If this is the case, English-speaking listeners would have to learn to direct their attention to onset f_0, separating it from a previously integral voicing dimension, to accurately identify Korean stop consonants.

Some clarification of our conceptualization of the role of attention in distinguishing phonetic contrasts is necessary. Just because an acoustic contrast is unattended does not mean that the acoustic differences are imperceptible to listeners or that such contrasts cannot be attended to in other contexts (including other speaking rates, the speech of other talkers, or other phonetic environments). Indeed, a distinction that is not attended to in one context may well be of crucial importance in another, whereas a contrast that is attended to in one context may be ignored in another. Although in principle any acoustic feature may be able to function as a cue (cf. Lindblom, 1990; Lisker, 1978), the mere availability of such cues need not imply that they will necessarily be used to make a particular phonetic decision. Experimental studies using conflicting cue patterns demonstrate that listeners show a clear hierarchy of preference to attend to particular cues over others, although this preference can change over the course of development or laboratory training (e.g., Francis, Baldwin, & Nusbaum, 2000; Nittrouer & Miller, 1997; Repp, 1982; Walley & Carrell, 1983). With limited attentional resources (Nusbaum & Schwab, 1986; Shiffrin & Schneider, 1977), it is expected that listeners will focus on those cues that have in the past proved to be most useful for identifying a particular contrast in a particular context (including phonetic context, speaking rate, and talker). Only those auditory features that have a high probability of being accurate predictors of a given linguistic contrast in a given context are likely to be attended to any significant degree. If auditory features covary reliably, listeners may process them together. If these features are attended together, listeners may treat them as a single integral dimension (see Smith & Kemler, 1977). Thus for cues that covary, such as VOT and f_0 in service of voicing decisions, English listeners may attentionally integrate these cues into a single perceptual dimension.

In some cases, learning a new contrast may simply involve learning to rely on the features of a contrast that, in prior experience, have not been found sufficiently distinctive (in terms of functional phonological contrast) to attend to separately. Indeed, it is interesting to note that listeners seem to have an easier time learning to hear unfamiliar foreign contrasts that are similar to acoustic contrasts present in their native language (e.g., the present study; McClaskey et al., 1983; Yamada & Tohkura, 1992) as compared with learning contrasts that they have never been exposed to (e.g., English speakers learning the Hindi retroflex–dental contrast; Tees & Werker, 1984) in a manner similar to the effect of preexposure on rats' learning of shape differentiation (Gibson & Walk, 1956). Thus, on the one hand, the fact that American listeners may be familiar with f_0- or CLEAR-based acoustic distinctions does not necessarily mean that they are attending to these as distinct cues, because these dimensions may not be as strongly predictive or as perceptually salient as the other cues to the voicing distinction in English with which they tend to covary, including VOT and amplitude of aspiration (see Lisker, 1978). One useful strategy for listeners in such a situation would be to incorporate weakly predictive cues into the perception of more strongly predictive cues with which they tend to covary, creating a complex, integral dimension. Whether listeners in this experiment are attending separately to f_0–CLEAR on the pretest is an empirical question. If no dimension in an MDS solution corre-

[6] We are grateful to an anonymous reviewer for pointing out most clearly the roles that f_0 and CLEAR might play in English.

lates with measured acoustic values of f_0–CLEAR, we have at least some support for the hypothesis that this dimension is not attended to as a distinct dimension of contrast. On the other hand, the likelihood of preexposure to onset f_0 differences that correlate with the phonological voicing contrast (as well as with variation in VOT that cues the same contrast) does suggest that English listeners will have a relatively easy time learning to attend to the f_0–CLEAR contrast in the laboratory if they do not already show evidence of attending to it on the pretest. The extraction of one component of an integral cue is conceptually distinct from the development of attention to a never-before encountered cue, and the distinction between these two processes may underlie differences in the ease of acquisition of different types of nonnative contrasts. Still, neither case is currently accommodated within existing A2D models, all of which assume that the set of possible dimensions is fixed in that they include no mechanism for developing new dimensions (either ex nihilo or by separation from preexisting integral dimensions; see Schyns, Goldstone, & Thibaut, 1998).

Method

Participants. Ten students from the University of Chicago (5 male and 5 female) participated in this experiment. All of the participants were native speakers of American English. All reported having normal hearing, and none had any experience hearing or speaking Korean. Because all prospective participants had some experience with at least one language other than English, preference was given to volunteers who had experience with only currently unspoken languages (Latin, classical Greek, American Sign Language). When participants had experience with a spoken foreign language, preference was given to those with little or no experience outside of high school or college classes. Volunteers who had lived abroad for a year or more, begun learning a foreign language before high school, or who spoke a language other than English on a regular basis were excluded from the study, though 1 participant who had begun learning French at age 11 was included accidentally. Although all participants reported at least some classroom experience with languages other than English, none of the languages reported has three classes of stop consonants.

Stimuli. Stimuli for this experiment were drawn from the same five sets of syllables described in the *Method* section of Experiment 1. In the present experiment, American participants were tested only on the syllables containing bilabial stops and the vowel /a/ from Lists 1 and 4 (both spoken in a conversational manner) for a total of six test syllables contrasting only in terms of the voicing quality of the stop consonant. For training, all other syllables were used. Thus, participants never heard any of the test syllables during training, and during training they were exposed to a variety of vowels (/a/, /o/, and /i/), places of articulation (bilabial, dental, and velar), and production styles (citation and conversational).

Because the training set contains syllables with the same syllable structure (CV) as the test set, spoken by the same talker, and in some cases even containing the same vowel /a/, we cannot test whether training has taught listeners to generalize from one talker (or phonetic context) to another. As is discussed below, generalization, or lack of it, is not the primary issue in this experiment. The purpose of using such similar training and test sets was to reduce the amount of training time necessary and improve the probability that listeners' categorization abilities would improve considerably, to ensure that the effects of training would be clearly discernible in the MDS solutions.

All stimuli were presented to participants binaurally at a comfortable listening level (approximately 65–75 peak dB SPL) over Sennheiser HD430 headphones in a sound-attenuated booth. Headphone level was under the control of each participant by means of a software interface, but few participants chose to change the level, and those who did change the level did not modify it beyond approximately ± 5 dB (as measured after the session in which level was adjusted). Presentation of the stimuli and collection of responses were digitally controlled on a SPARC workstation using a software interface.

Procedure. Participants took part in three sessions, the first and last of which took approximately 60 min, with the second requiring about 40 min. Participants were paid $35 at the end of the experiment. As shown in Table 3, the first and last sessions consisted primarily of the pretest and posttest phases, whereas the middle session consisted entirely of training. Participants also received some training at the start of the third session, immediately preceding the posttest. During the first test session, participants were first given a description of the entire experiment and then completed two pretest tasks: a perceptual difference rating (inverse similarity) task and a phonetic identification task. In the posttest, participants repeated the same tasks in reverse order.

The identification task consisted of 10 presentations of each of the six syllables (two tokens for each of three consonant classes [pa], [p^ha], and [Pa]), in random order. Participants were instructed to respond by clicking on one of three buttons labeled with pseudo-phonetic transcriptions of the three consonants ([p^h] was written as *ph*, [P] as *pp*, and [p]

Table 3
Training Experiment Procedure: Schedule and Major Characteristics of Experimental Blocks

Day	Session	Task	Block	Trials	Response
1	Pretest	Difference rating	Familiarization	1	None
			Testing	144	Slider scale rating (0–100)
		Identification	Familiarization	2	None
			Testing	60	3 AFC (p, ph, pp)
2	Training	3 blocks	Familiarization	2	None
			Training	129 per block (387 total)	9 AFC (p, ph, pp, t, tt, th, k, kk, kh)
3	Training	1 block	Familiarization	2	None
			Training	129	9 AFC (p, ph, pp, t, tt, th, k, kk, kh)
	Posttest	Identification	Familiarization	2	None
			Testing	60	3 AFC (p, ph, pp)
		Difference rating	Familiarization	1	None
			Testing	144	Slider scale rating (0–100)

Note. AFC = Alternative forced-choice task.

as *p*). Before beginning the test, during familiarization, participants heard two instances of one good prototype ([pa], [p^ha], and [Pa] from List 3, produced in citation form) of each of the three categories and were shown which symbol corresponded to each sound without making a response. On the identification task, responses were scored as correct if the selected symbol corresponded to the category from which the stimulus was selected.

The difference rating task contained two parts. The first part was to give participants an idea of the overall range of variation between the syllables in this task, and it provided one auditory presentation of every test syllable ([pa], [p^ha], [Pa] from Lists 1 and 4, produced in conversational style) at a rate of approximately one token per second. In the second part, participants were given 144 difference-rating trials (4 trials with each of the 36 pairs in the difference-rating set. That is, all pairwise combinations of [pa], [pha], and [Pa] from Lists 1 and 4) with 250 ms interstimulus interval for each pair. Participants were instructed to rate the degree of difference (if any) between each pair of sounds by setting a slider bar on a computer screen. In each trial, the slider on the bar appeared at the far left of the scale. No numbers were displayed, but the output response of the scale ranged from 0 (labeled *identical*) on the left to 100 (no label) on the right. Participants were instructed to think of the scale as "extending from identical (no difference) at the left to 100% different—that is, as different as any two consonants in the set could possibly be at the right." The trough of the slider was approximately 10 cm long. Each pair of the six CV syllables was presented four times in each test, for a total of 144 ratings during the pretest and 144 ratings during the posttest.

Beginning on the 2nd day of the experiment, participants were trained to recognize exemplars from all three voicing categories. The training phase of the experiment consisted of four presentations of 129 training syllables (Lists 2, 3, and 5, each consisting of 27 syllables plus Lists 1 and 4 excluding the /p^ha/, /Pa/, and /pa/ syllables, which were reserved for testing). On each training trial, participants were asked to identify the syllable they heard by clicking on the button marked with the appropriate pseudo-phonetic symbol. Transcription conventions during training followed those in the identification task. Thus /p^h/ was written as *ph*, /t^h/ as *th*, /k^h/ as *kh*, /p/ as *p*, /t/ as *t*, /k/ as *k*, /P/ as *pp*, /T/ as *tt*, and /K/ as *kk*. Note that listeners were trained with syllables containing consonants at all three places of articulation (bilabial, alveolar, and velar) to facilitate learning, but they were tested using only the bilabial stimuli excluded from the training set. As in the identification session, participants heard two instances of a good exemplar (in citation form, from List 3) of each of these nine categories prior to starting training (these exemplars were also included in the training set).

During training, if participants identified a consonant incorrectly, they were shown the correct symbol and heard a repetition of the stimulus. They were not given a chance to correct their selection. If participants clicked on the correct symbol, they were informed that they were correct and heard the stimulus again as reinforcement. Participants performed this task four times on the complete list of 129 syllables, in random order. The first three repetitions of the list were done on the 2nd day of the experiment, whereas the fourth repetition was done on the 3rd day of the experiment, immediately prior to beginning the posttest.

Results

Learning. Consonant identification scores improved by 33 percentage points, from a mean score of 53% correct on the pretest to 86% correct on the posttest (where chance is assumed to be 33% correct on both tests). This improvement was significant, $t(9) = 3.97$, $p < .01$.[7] Participants were noticeably above chance even on the pretest, reflecting their generally good discrimination of the strong consonants. To determine whether training affected the perceived similarity of stimuli, we grouped responses according to stimulus pairs. Same-category pairs include pairs of different utterances of the same category (produced in different recording lists, e.g., $[p^h]_4$–$[p^h]_1$ and $[p^h]_1$–$[p^h]_4$) as well as pairs of identical tokens (e.g., $[p^h]_4$–$[p^h]_4$). Different-category pairs include all pairs in which the two tokens are from different linguistic categories (e.g., $[p^h]_1$–$[P]_1$ and $[p^h]_1$–$[P]_4$). It was expected that training would encourage participants to treat same-category pairs as more similar and between-categories pairs as less similar to improve categorical perception of the stimuli (Livingston et al., 1998; Studdert-Kennedy, Liberman, Harris, & Cooper, 1970). As shown in Figure 4, this assumption is only partially supported. When the average difference scores for each pair of tokens are examined, we see a main effect of category (same vs. different), $F(1, 34) = 88.25$, $p < .01$, and of test (pretest vs. posttest), $F(1, 34) = 12.31$, $p < .01$, but no interaction, $F(1, 34) = 2.91$, *ns*. Different-category pairs increased in difference by an average of 8 points, from 57.2 to 65.2, and this difference was significant according to a planned comparison of means, $F(1, 34) = 19.05$, $p < .01$.[8] However, contrary to prediction, same-category pairs also increased very slightly in difference by an average of 3 points, from 7.2 to 10.2, though this difference was not significant by planned comparison, $F(1, 34) = 0.61$, $p = .44$. Examination of the difference ratings of individual pairs of consonants reveals the following:

1. Looking only at the pairs containing a /P/ token, the average difference of different-category pairs increased, as predicted, from 72.1 to 75.9, which is significant by planned comparison of means, $F(1, 18) = 8.04$, $p = .01$. Meanwhile, the average difference of same-category pairs decreased, as predicted, from 5.8 to 3.9, but this change is not significant, $F(1, 18) = 1.72$, *ns*. However, this nonsignificant result is due to the inclusion of pairs of identical tokens ($[P]_1$–$[P]_1$ and $[P]_4$–$[P]_4$) that already have a mean rating of almost zero on the pretest (0.588) and drop completely to zero on the posttest. Excluding these pairs from the analysis shows that the decrease in mean difference ratings of different pairs containing a /P/ is significant, $F(1, 16) = 12.79$, $p < .01$. This pattern of results suggests that listeners have learned to perceive /P/ tokens as more similar to one another and as more different from other tokens as a result of training.

2. For the pairs containing a /p^h/ token, the average difference of different-category pairs increased, as predicted, from 50.6 to 59.4, and this difference is significant, $F(1, 18) = 13.33$, $p < .01$. Meanwhile, the average difference of same-category pairs also increased, contrary to prediction, from 6.2 to 14.8, and this difference is also significant, $F(1, 18) = 5.43$, $p = .03$. This pattern of results suggests that listeners have learned to treat /p^h/ tokens as more different from other tokens but also as less similar to one another.

3. For the pairs containing a /p/ token, the average difference of different-category pairs increased, as predicted, from 48.9 to 60.3, and this change is significant, $F(1, 18) = 43.62$, $p < .01$. Meanwhile, the average difference of same-category pairs also increased slightly from 9.6 to 11.9, but this change is not significant, $F(1,$

[7] All tests using residual mean squares comparing proportions are based on arcsine-transformed percentages to ensure that block and treatment effects are additive (Kirk, 1995).

[8] Mean difference ratings on a scale of 0–100 were converted to percentages (0–1) prior to application of the arcsine transformation and subsequent statistical analyses.

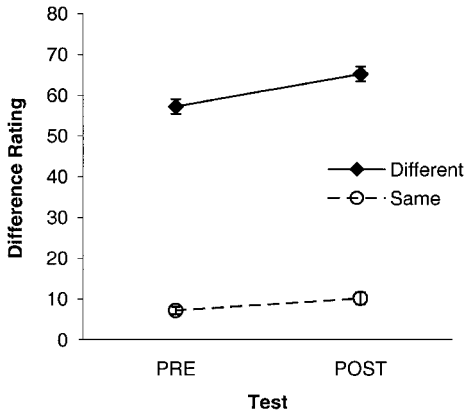

Figure 4. Change in difference ratings of Korean consonant pairs from pretest to posttest (Experiment 3), showing differential change in ratings for pairs in which both tokens are from the same phonetic category (dotted line) and for pairs in which both tokens are from different categories (solid line). Error bars indicated standard error.

18) = 0.06, *ns*. This pattern of results suggests that listeners have learned to treat /p/ tokens as more different from other tokens, but probably not as more different from one another.

These results suggest that listeners learned to treat different-category pairs as more different overall, but their evaluation of same-category pairs varied depending on the category. To identify changes in the weighting of individual dimensions of contrast, we submitted participants' difference ratings to an MDS analysis.

Multidimensional scaling. A series of solutions in various numbers of dimensions was calculated to determine the optimal dimensionality for the final scaling. As shown in Figure 5, the fit values for the pretest and posttest solutions are similar across all measured dimensionalities above 1, but they are noticeably different for solutions of one dimension, suggesting that training has changed the optimal dimensional structure of listeners' perceptual space.

Looking only at the fit values for a one-dimensional solution, a one-dimensional solution for listeners' pretest difference ratings already has a very good fit correlation (.987). In contrast, the one-dimensional solution for listeners' posttest difference ratings has somewhat poorer goodness of fit (.958). In other words, one effect of training seems to be to reduce the effectiveness of a one-dimensional solution. This suggests that while a one-dimensional solution is likely to be most appropriate for the pretest data (solution plotted in Figure 6), the posttest data require at least a two-dimensional solution (plotted in Figure 7).

To identify the dimensions to which listeners were attending on each test, we compared the dimensional coordinates of each token in the MDS solution with the measured values of four acoustic parameters, as discussed in the introduction. Note that in the present stimuli, measured values of VOT and RISE are correlated ($r = .62$, $p = .01$), as are CLEAR and f_0 ($r = .60$, $p = .01$). The presence of such multicollinearity between independent variables artificially increases standard errors in multiple regression analyses, reducing the precision of estimated regression coefficients (Schroeder, Sjoquist, & Stephan, 1986). Therefore, although multiple regression analysis is typically used to distinguish the relative weight of the dimensions derived using MDS analyses (Kruskal & Wish, 1978; Livingston et al., 1998), it was not applied in this case, and separate regression analyses were carried out for each dimension and acoustic parameter. The results of these analyses are shown in Table 4.

Although the solution space is not plotted here, a two-dimensional solution was calculated for the pretest to determine whether a higher dimensional analysis of the pretest might show evidence that listeners were attending to the learned dimensions even prior to training. However, Table 4 clearly shows that, on the pretest, listeners appear to be attending to VOT or RISE or both, but not at all to CLEAR or f_0 regardless of the number of dimensions in the scaling solution. It is important to note that whether we adopt the one- or the two-dimensional pretest solution, it is clear from Table 4 that, on the posttest, listeners treated f_0 (or CLEAR) as an acoustic feature that can be attended to separately from VOT, yet there is no evidence that they were doing this on the pretest. On the basis of the correlations shown in Table 4, it also appears that listeners may be attending less to the VOT–RISE dimension on the posttest than they are on the pretest. This may be an indication that attentional resources are limited, and shifting attention to a new cue (or one separated out of the old composite voicing [VOT] cue) may require the reduction of the amount of attention directed to other cues.

The solutions derived by the MDS procedure are normalized, which precludes a direct comparison of raw distances between points in one solution space with those in another. However, because ratios of distances are not affected by normalization, it is possible to compare a ratio of distances along a single dimension between two pairs of points in the pretest solution with that of the same points in the posttest. If this ratio changes, that would provide strong evidence that the dimension has stretched or shrunk nonuniformly as a consequence of training, contradicting the predictions of the GCM but not the localized stretching model. Indeed, evidence for such nonuniform changes is quite clear in comparing the relative distances of points ordered along the VOT–RISE dimension in Figure 4 with those in Figure 5. For example, the ratio $[p^h]_1$–$[p^h]_4/[p^h]_1$–$[P]_1$ is .14 on the pretest but .005 on the posttest, indicating that, along the VOT–RISE dimension, the perceived difference between points $[p^h]_1$ and $[p^h]_4$ has decreased

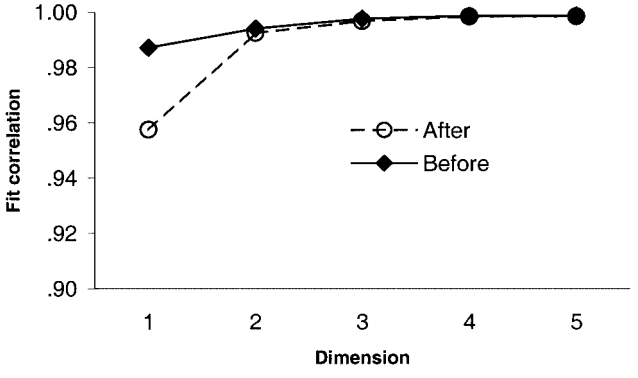

Figure 5. Fit correlation by dimensionality for trained participants' pretest and posttest difference ratings (Experiment 3). The solid line with solid symbols shows fit correlation values for the pretest, whereas the dashed line with hollow symbols shows fit correlation values for the posttest.

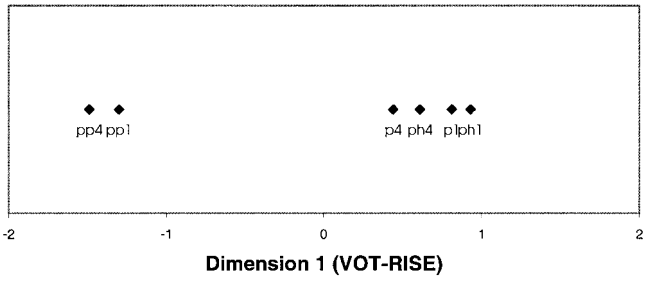

Figure 6. One-dimensional pretest solution for trained listeners (Experiment 3). Tokens are transcribed as in Han and Weitzman (1970), with the exception that /P/ is written here as /pp/ and /p^h/ as /ph/. Numerals refer to the recitation list that the token is drawn from (see *Method* section of Experiment 1). VOT = voice onset time; RISE = measured duration from onset of vowel formants to peak vowel amplitude.

by more than has the perceived difference between points $[p^h]_1$ and $[P]_1$.

To compare ratios of distances more systematically, it is possible to calculate a *structure ratio*, defined as the ratio of within-category distances to between-categories distances (Cohen & Segalowitz, 1990). This structure ratio can then be used to investigate the relative changes in the distribution of attention within and between categories. Looking just at the changes in structure ratios involving the first dimension, VOT–RISE, the average ratio of distances within the category /p/ along this dimension to those between tokens in /p/ and tokens outside of /p/ along this dimension increases from 1.08 to 11.66, indicating that the within-category distances increased considerably more than did the between-categories distances. For /p^h/ the change was very much the opposite, from .92 to .19, whereas for /P/ there was hardly any change at all, from .09 to .10. These results suggest that listeners learned to treat /p/ tokens as more different according to VOT–RISE after training while simultaneously learning to treat /p^h/ tokens as *less* different along the same dimension.

It is also possible to compare the perceptual spaces of different groups of participants in terms of their degree of correlation with the measured acoustic space, as shown in Table 5. From this table, it can be seen that trained participants have learned to attend to the dimensions of phonetic contrast in a manner approximating that of native speakers. This is in contrast to the untrained (pretest) pattern of correlations. Comparing Figure 7 with Figure 3 (showing native Korean speakers' perceptual space) provides further support for this observation. The relative location of trained participants' categories (as determined by the locations of the two-member tokens) is approximately the same as that of native speakers. If we compare the perceptual space derived for native speakers in Experiment 2 with that derived for English speakers' posttest results, there is a strong correlation between the token locations derived from native results and those derived from trained listeners' posttest results (native [Dimension 1] vs. trained posttest [Dimension 1], $r = .98$, $p < .01$; native [Dimension 1] vs. trained posttest [Dimension 2], $r = .19$, $p = .72$; native [Dimension 2] vs. trained posttest [Dimension 1], $r = .01$, $p = .99$; and native [Dimension 2] vs. trained posttest [Dimension 2], $r = .90$, $p = .01$).

The observation that there is a high degree of correlation between the distribution of stimuli in trained listeners' perceptual space in the present experiment and the distribution of the same tokens in native listeners' perceptual space as derived in Experiment 2 strongly supports the claim that trained participants were indeed learning to attend more strongly to the same dimensions that native speakers use. An analysis of the correlation between various acoustic parameters (identified and measured in Experiment 1) with both trained participants' results (derived here) and native participants' results (derived in Experiment 2), as shown in Table 5, further suggests that trained listeners have indeed learned to attend to the same dimensions of contrast attended to by native speakers of Korean.

Discussion

The present results suggest that, in learning a new phonetic contrast, listeners may restructure their perceptual space both by warping existing dimensions in a localized manner and by changing the distribution of attention to divide previously integral auditory dimensions (e.g., Schyns et al., 1998) or to attend to properties that were not previously used for phonetic contrasts. The results of the analysis of difference ratings presented here further suggest that perceptual learning can involve both increased within-category similarity and increased between-categories differences, in a manner similar to Kuhl and Iverson's (1995) account of the acquisition of native-language phonetic categories. Similarly, the results of the analysis of structure ratios calculated along the VOT–RISE dimension suggest that the application of compression

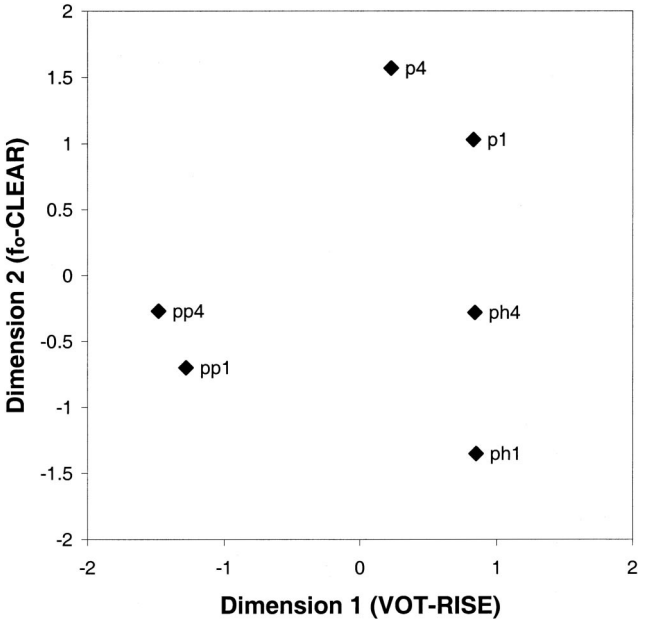

Figure 7. Two-dimensional posttest solution for trained listeners (Experiment 3). Tokens are transcribed as in Han and Weitzman (1970), with the exception that /P/ is written here as /pp/ and /p^h/ as /ph/. Numerals refer to the recitation list that the token is drawn from (see *Method* section of Experiment 1). CLEAR = clarity of formant structure at onset of phonation; f_0 = measured fundamental frequency; VOT = voice onset time; RISE = measured duration from onset of vowel formants to peak vowel amplitude.

Table 4
Correlations and p Values of Stimulus Parameter Values With Locations of Bilabial Tokens in Various Solution Spaces Calculated for Trained Participants

	1-D pretest scaling, Dimension 1		2-D pretest scaling, Dimension 1		2-D pretest scaling, Dimension 2		2-D posttest scaling, Dimension 1		2-D posttest scaling, Dimension 2	
Parameter	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
VOT	*.77*	*.071*	.70	.125	**.96**	**.003**	*.79*	*.060*	−.41	.420
RISE	**.91**	**.011**	**.93**	**.007**	.44	.382	**.86**	**.029**	.41	.420
f_0	−.35	.499	−.46	.364	.41	.421	−.29	.580	**−.89**	**.017**
CLEAR	−.02	.966	−.14	.796	.70	.123	.06	.909	**−.91**	**.013**

Note. Correlations significant at or below the $p = .05$ level are marked in bold. Nearly significant correlations ($p < .10$) are in italics. Measured values for all parameters of all tokens are shown in Table 1. VOT = voice onset time; RISE = measured duration from onset of vowel formant to peak vowel amplitude; f_0 = measured fundamental frequency; CLEAR = clarity of formant structure at onset of phonation.

or expansion is a function of the relationship between categories prior to training.

In the case of the category /P/, training resulted in compression within the category and expansion between its members and those of other categories, as shown in the changes in difference ratings. The lack of a change in structure ratios involving tokens from /P/ along the VOT–RISE dimension suggests that any changes within this category must have involved other dimensions. The changes observed in the other two categories suggest that one significant consequence of training is the development of a new dimension of contrast, f_0–CLEAR. In the case of the category /p^h/, the observed, and unexpected, increase in difference ratings within the category can clearly be explained by the development of attention to this new dimension. Members of /p^h/ become perceived as more similar to one another than to members of other categories along the VOT–RISE dimension, but they end up being treated as more different overall because of their differences along the f_0–CLEAR dimension. Changes in the /p/ category, in contrast, are most likely due primarily to changes in the VOT–RISE dimension, along which they become increasingly more different (in comparison with their perceived similarity to tokens from other categories) as a result of training. This increased difference along one dimension is not, however, sufficient to override the contribution to difference ratings of the increased attention to f_0–CLEAR along which these tokens differ even more from members of the other categories.

The analyses of changes in difference ratings of individual pairs of stimuli further suggest that, in this case, acquired distinctiveness between categories results partly from attending to a dimension of contrast in the posttest that does not appear to have played a noticeable role in listeners' pretest difference ratings. Tokens from the /P/ category are already easily distinguished on the basis of VOT (or RISE) alone and are not strongly affected by the increased attention to f_0 (or CLEAR). This may be because the two tokens in this category do not differ much along this dimension, or possibly because this dimension is simply not relevant for distinguishing members of this category from other tokens (see the General Discussion below). In contrast, it is necessary for listeners to attend to an additional dimension of phonetic contrast to distinguish between tokens from the /p/ and /p^h/ categories, even though doing so increases within-category differences as well.

However, it is not clear from the present results whether this restructuring would hold for different places of articulation or different talkers. Although listeners were never exposed to the specific test stimuli during training, all of the stimuli were produced by the same talker and had the same syllable structure (CV). In addition, a subset of the training stimuli even contained the same vowel (/a/) as the test stimuli. Thus, although it is certainly the case that listeners learned more than a simple association of linguistic labels with particular acoustic patterns (as they were never trained on any of the test stimuli), the results presented here are not sufficient to support the strong claim that listeners have

Table 5
Correlations and p Values of Stimulus Parameter Values With Locations of Tokens in Various Solution Spaces

	Native, Dimension 1		Native, Dimension 2		Trained pretest, Dimension 1		Trained posttest, Dimension 1		Trained posttest, Dimension 2	
Parameter	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
VOT	*.78*	*.065*	−.52	.287	*.77*	*.071*	*.79*	*.060*	−.41	.420
RISE	**.84**	**.040**	.30	.562	**.91**	**.012**	**.86**	**.028**	.41	.420
f_0	−.21	.692	**−.95**	**.004**	−.35	.499	−.29	.580	**−.89**	**.017**
CLEAR	.06	.905	**−.92**	**.010**	−.02	.966	.06	.909	**−.91**	**.017**

Note. Correlations significant at or below the $p = .05$ level are marked in bold. Nearly significant correlations ($p < .10$) are in italics. Stimulus values for all parameters for all tokens are shown in Table 1. VOT = voice onset time; RISE = measured duration from onset of vowel formant to peak vowel amplitude; f_0 = measured fundamental frequency; CLEAR = clarity of formant structure at onset of phonation.

learned to identify a generalized rate-, talker-, or context-independent linguistic feature used in Korean to distinguish three categories of voicing. At best, we can state only that listeners show evidence of shifting the distribution of their attention along existing dimensions in a manner consistent with localized warping A2D models, and that they appear to be attending to an acoustic dimension in the posttest that they showed no evidence of attending to in the pretest. However, these results may be specific to /Ca/ syllables produced by this particular talker.

These results are consistent with Livingston et al.'s (1998) claim that compression and expansion are used to increase the linear separability of categories in perceptual space. In the case of Goldstone's (1994) studies in which tokens close to category boundaries were within one just-noticeable difference (JND) of one another, expansion at the boundaries (enhancing discriminability) is necessary for separating categories. In Livingston et al.'s experiments, compression within categories was sufficient to enable linear separability, and no expansion at category boundaries was necessary. Livingston et al. suggested that this may be due to the readily identifiable differences between categories in their studies, even before training. They argued that it may be more efficient for participants in their experiments to learn to ignore distinctions within categories than it would be to learn to attend even more strongly to distinctions between two already distinguishable categories. The results presented here support this hypothesis. As in Livingston et al.'s experiments, some of the stimuli used here (the /P/ tokens) are relatively easily distinguished from the others prior to training, and therefore compression within this category is more efficient than expansion between it and other categories. However, compression within the /P/ category does not increase the discriminability of the /p/ category from the /p^h/ category, and might even reduce it. Even the apparent (local) expansion of VOT–RISE, observable in the changes in the VOT–RISE structure ratios for /p/ tokens, does not appear sufficient to distinguish /p/ from /p^h/. To separate /p/ from /p^h/ requires the inclusion of another dimension of contrast. In this case, the necessary dimension, f_0–CLEAR, is likely available to listeners as a component of an integral voicing dimension with which they are familiar from English. In this experiment, listeners appear to have shifted the distribution of their attention to increase the linear separability of all three categories, locally compressing dimensions to decrease the distance between members of well-defined categories, expanding familiar dimensions in regions between two poorly distinguished categories, and learning to attend separately to the formerly integral components of a familiar dimension.

General Discussion

The phonetic contrast investigated in the present experiments has not been the subject of a nonnative perceptual training study until now. Thus our results contribute to the body of research suggesting that participants can learn a new phonetic contrast given the appropriate training (e.g., Jamieson & Morosan, 1986; Logan, Lively, & Pisoni, 1991, McClaskey et al., 1983; Tees & Werker, 1984), although the present study does not constitute a strong case of phonetic learning because we did not attempt to assess the degree of generalization to new talkers or contexts. However, in previous second-language phonetic training studies in which such learning has been demonstrated, there is typically also generalization when it is tested (e.g., Logan et al., 1991; McClaskey et al., 1983). The question of whether listeners have learned to attend to an acoustic contrast that is generally useful in a talker-, rate-, and phonetic context-independent manner is important, but because of the limitations of stimulus design, it cannot be the primary focus of the present analysis. In this article, we can only focus on the question of whether listeners exhibited certain predicted changes in spatial representations of their perception of these stimuli, reflecting a shift of attention to a new dimension as a result of identification training. This question is important for understanding how learning changes the distribution of attention during phonetic classification and whether these changes are consistent with the predictions of A2D models of categorization. Whether listeners show these kinds of changes in circumstances of more general phonetic learning remains to be determined.

Indeed, it is possible that listeners naturally shift their attention between cues as they shift between talkers, phonetic contexts, and speaking rate as part of the process of context normalization. For example, it may well be the case that different talkers produce different relative weightings of acoustic cues to a particular contrast—a cue that is highly predictive of a particular contrast in the speech of one talker may be less clearly useful in the speech of another talker—and therefore listeners will do better to shift their distribution of attention between these cues as they change from one talker to another. This hypothesis is supported by the observation that talker normalization is a resource-demanding process (Nusbaum & Magnuson, 1997; Nusbaum & Morin, 1992), suggesting that changing talkers requires listeners to change something about the way they listen. Strong tests of learning involving cross-talker or cross-context generalization may therefore reflect only the acquisition of the most robust (context-independent) acoustic cues, not necessarily all the cues that listeners use in a particular context.

Despite these limitations on generalization, our results provide some insight into three aspects of the application of A2D dimensional warping models of perceptual learning to the study of phonetic category acquisition. First, the results of training are consistent with a model of perceptual learning in which improvement at distinguishing between categories results both from a process of acquired similarity within categories and from a process of acquired distinctiveness between categories. This pattern of results also suggests a resolution to the apparent disagreement between Goldstone's (1994) observations and those discussed by Livingston et al. (1998). When categories are sufficiently differentiable before training, training may serve primarily to increase within-category similarity. However, when categories are indistinguishable (e.g., tokens on either side of the category boundary lie within a single JND), listeners increase their attention to those dimensions that differentiate the two categories, expanding the distinctions between them.

To increase distinctiveness between categories, listeners may also have to attend in new ways to stimuli. Training to recognize a foreign phonetic contrast can apparently cause listeners to change the distribution of attention that maps auditory properties onto phonetic categories, at least in cases in which two categories are not linearly separable along existing dimensions of contrast. This observation is consistent with Livingston et al.'s (1998) observation that perceptual learning can involve a change in the dimensional structure of perceptual space, but it cannot be ac-

counted for in terms of current formulations of existing dimensional warping theories of perceptual learning. It is important to distinguish between the development of a completely new dimension of phonetic contrast and the distribution of attention to a dimension separated out of a more complex (integral) dimension. Although these two processes are in principle distinct, they are both equally poorly accommodated within current A2D models of perceptual learning (see Schyns et al., 1998, for a more in-depth discussion of the development of new features in the service of visual categorization).

One way of reconciling these observations with current A2D dimensional warping models is to argue that listeners did not actually develop a new contrast (or even pull a new dimension out of an old one). Rather, they simply learned to direct more attention over the entire length of an already existing dimension of contrast that was initially attended to only to a vanishingly small degree, in effect stretching the dimension out from a length very close to zero (a possibility discussed by Schyns et al., 1998). This solution is consistent with the observation that English speakers have been exposed to the covariation of onset f_0 and stop consonant class. According to this hypothesis, the listeners in the present experiments showed no evidence of attending to onset f_0 as a separate dimension on the pretest because they were used to ignoring (or discounting) it in English, in favor of attending to other, stronger cues such as VOT. This discounted dimension was not absent from their perceptual space (nor was it integrated into some other, more complex dimension), but rather it merely atrophied from disuse to such a degree that the small amount of residual attention directed to it was not detectable with the techniques used here. When listeners discovered that this dimension was in fact necessary for the present experiment, they simply (re)directed attention to it and stretched it out again.

Although this hypothesis is consistent with the results presented here, it may not be sufficient to account for other cases of perceptual learning such as those in which the foreign contrast is clearly not present in any way in the listeners' native language or prior experience (e.g., the Hindi retroflex–dental stop contrast for English listeners). This suggests that one possible problem with the stretching from zero argument is that it requires one to accept that the set of possible phonetic dimensions is innate, finite, and probably relatively small, because it presumes that all cases of phonetic learning involve at most redirected attention to a previously ignored dimension; this model rejects a priori the possibility of learning a truly novel dimension of phonetic contrast. However, there is no clear evidence that there are any limitations on the acoustic differences that could constitute a learnable dimension of phonetic contrast, other than those imposed by the physiology of the human auditory system (Lindblom, 1990). Although it would be impossible to prove that there are no such limitations, it is well known that listeners are able to perceive a wide range of sounds as speech, including articulatory exotica such as the click phonemes of Khoisan and some southern Bantu languages that are typically not perceived as speech by nonnative listeners (see Best et al., 1988). Besides clicks, lesser known examples of nonspeech sounds being used linguistically include whistled speech and talking drums (Sebeok & Umiker-Sebeok, 1976), as well as sine-wave speech (Remez, Rubin, Pisoni, & Carrell, 1981) and the vocalizations of parrots and other avian mimics (Pepperberg & Neapolitan, 1988). These examples demonstrate that human beings are quite capable of learning to use a phenomenal range of acoustic phenomena in a phonetically contrastive manner, suggesting that, at the very least, the set of possible contrastive dimensions is very large. As attention is generally considered a capacity-limited resource (see Nusbaum & Schwab, 1986), it does not seem any more reasonable to assume, a priori, that listeners are always attending to a potentially infinite number of dimensions of phonetic contrast than to assume that it is possible to learn a completely new dimension of phonetic contrast.

Similar arguments may be used to suggest that not all cases of perceptual learning can be accounted for in terms of the separation of previously integral dimensions of contrast. After all, it may merely be fortuitous that the Korean stop contrast is made along an acoustic dimension that covaries with, and can be separated out of, a voicing dimension in English. How would a separation and integration model of phonetic learning account for cases such as the Hindi retroflex–dental contrast for English listeners? For this, it is necessary to consider phonetic features as psychological constructs rather than perceptual primitives. That is, it may be more useful to consider phonetic features such as voicing or manner in terms of complex patterns of auditory (neural) features extracted by the nervous system, rather than in terms of atomic acoustic–phonetic features measured from the waveform. Indeed, the fact that even relatively simple acoustic patterns are encoded in multiple, parallel neural systems provides some tentative support for this hypothesis (Steinschneider, Arezzo, & Vaughan, 1982; Sussman, 1988). Treating phonetic features in auditory terms would make it possible to consider the development of new phonetic contrasts in terms of chunking existing auditory features of contrast and, if necessary, feature decomposition of high-level phonetic features that lose their relevance in a particular categorization task. In a sense, the primary difference between the stretching-from-zero and the separation-and-integration models proposed here lies in the level at which the features of speech perception are presumed to be nondecomposable. In the first case, acoustic–phonetic features (or their auditory signature) such as VOT or onset f_0 are considered fundamental, whereas in the second case the basic components of phoneme perception are presumed to be simple patterns of low-level neural activity (which combine to form the neural patterns related to traditional acoustic–phonetic features). This is obviously an area of research far beyond the scope of this article. However, it may be noted that only the second model treats phonetic features as flexible in a manner consistent with Schyns et al.'s (1998) arguments in favor of the existence of flexible feature systems.

The results of training also suggest that shifting attention in a context-dependent manner can affect learning, which is consistent with the hypothesis that the set of features attended to is (in part) a function of the categorization tasks that must be performed (Schyns et al., 1998). Recall that trained participants generally increased the perceived overall difference within the /p^h/ and /p/ categories, but /P/ tokens were not as strongly affected. One possible reason for this is that listeners may not have directed attention to f_0 or CLEAR to the same degree for all pairs. This dimension plays little or no role in distinguishing /P/ tokens from those in the other two categories because VOT (or RISE) is already sufficient to make the necessary categorical distinction. In making distinctions that include a /P/, listeners did not need to attend to f_0 or CLEAR to the same degree as they do for pairs that include only

tokens from the other two categories. Thus, it is possible that listeners only learned to use f_0 or CLEAR in those cases in which they experienced a benefit from attending to it when making identification judgments during training. Alternatively, this could reflect a regional warping of perceptual space—a stretching of the f_0–CLEAR dimension only at higher values of VOT–RISE. This kind of regional warping is not commensurate with a uniform dimensional warping model like the GCM but could be modeled in a localized model such as that described in Goldstone (1994) or Guenther et al.'s (1999) connectionist model.

It is also possible that the observed context-dependent shifting of attentional focus is an artifact of the difference-rating task rather than a characteristic of normal speech perception processes. That is, asking participants to overtly rate the similarity or difference of a pair of tokens may artificially encourage them to focus attention on those features along which those tokens differ most significantly or consistently in a manner or to a degree that is not present in normal speech perception. However, the observation that listeners are able to shift the focus of their attention in a context-dependent manner is consistent with a number of theories of speech perception. In particular, the hypothesis-testing mechanism proposed by Nusbaum and Schwab (1986) and Nusbaum and Magnuson (1997) predicts specifically that listeners will attend most strongly to those acoustic dimensions that are most reliable for distinguishing between those linguistic categories that are possible in a given context. According to this view, learning a new phonetic contrast is not simply a matter of retuning perceptual filters but rather involves learning which acoustic features to attend to in which contexts (Francis et al., 2000; Nusbaum & Goodman, 1994).

Conclusions

To apply current selective attentional models of perceptual learning to account for phonetic learning, one must determine whether the process of phonetic learning in fact conforms to the predictions made by these models. The results of the experiments described here provide some support for the application of A2D models to phonetic categorization, but they also provide some preliminary evidence that these models may be insufficient as currently formulated. The incorporation of models of selective attention into speech perception research represent a step in the right direction for accounting for phonetic learning because these models provide a unified account of phonetic learning in terms of changes in the distribution of attention—a concept that is pervasive in the phonetic learning literature. In particular, the results presented here corroborate the findings of Iverson and Kuhl (Iverson & Kuhl, 1995, 2000; Kuhl & Iverson, 1995) that phonetic learning can involve both acquired similarity and acquired distinctiveness along the same dimension, suggesting that localized warping models may be more successful than models such as the GCM in accounting for the facts of phonetic learning. However, some modification to the existing models may be necessary to account for the development of new dimensions of contrast, especially if the intuitive stretching-from-zero argument can be shown not to hold in some future case of the acquisition of foreign phonetic contrasts that have no analogue in listeners' native language. It is possible that recent work on flexible feature systems may combine usefully with an auditory–neural perspective on phonetic features to further the development of more comprehensive A2D models of perceptual learning.

References

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In H. C. Nusbaum & J. Goodman (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167–224). Cambridge, MA: MIT Press.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). Baltimore: York Press.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 345–360.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics, 61,* 977–985.

Burnham, D. (1986). Developmental loss of speech perception: Exposure to and experience with a first language. *Applied Psycholinguistics, 7,* 207–240.

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

Fox, R. A. (1985). Multidimensional scaling and perceptual features: Evidence of stimulus processing or memory prototypes? *Journal of Phonetics, 13,* 205–217.

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

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

Gibson, E. J., & Walk, R. D. (1956). The effect of prolonged exposure to visually presented patterns on learning to discriminate them. *Journal of Comparative and Physiological Psychology, 49,* 239–242.

Goldstone, R. (1993). Feature distribution and biased estimation of visual displays. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 564–579.

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

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

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

Haggard, M., Ambler, S., & Callow, M. (1970). Pitch as a voicing cue. *Journal of the Acoustical Society of America, 47,* 613–617.

Han, M. S., & Weitzman, R. S. (1965). *Studies in the phonology of Asian languages: Vol. 3. Acoustic characteristics of Korean stop consonants.* Los Angeles: Acoustic Phonetics Research Laboratory, University of California.

Han, M. S., & Weitzman, R. S. (1970). Acoustic features of Korean /P, T, K/, /p, t, k/ and /p^h, t^h, k^h/. *Phonetica, 22,* 112–128.

Hardcastle, W. H. (1973). Some observations on the *tense–lax* distinction in initial stops in Korean. *Journal of Phonetics, 1,* 263–272.

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

Iverson, P., & Kuhl, P. (1996). Influences of phonetic identification and

category goodness on American listeners' perception of /r/ and /l/. *Journal of the Acoustical Society of America, 99,* 1130–1140.

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.

Jamieson, D. G., & Morosan, D. E. (1986). Training non-native speech contrasts in adults: Acquisition of the English /θ/–/ð/ contrast by francophones. *Perception & Psychophysics, 40,* 205–215.

Jusczyk, P. (1994). Infant speech perception and the development of the mental lexicon. In H. C. Nusbaum & J. Goodman (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 227–270). Cambridge, MA: MIT Press.

Jusczyk, P. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.

Kang, K.-S. (1998). On the phonetic parameters in the acquisition of Korean obstruents: A case study. In M. C. Gruber, D. Higgins, K. S. Olson, & T. Wysocki (Eds.), *Proceedings of the 34th annual meeting of the Chicago Linguistic Society: Vol. 2. The panels* (pp. 311–326). Chicago: Chicago Linguistic Society.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22–44.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling* (Sage University Paper Series on Quantitative Applications in the Social Sciences, No. 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.

Lindblom, B. (1990). On the notion of "possible speech sound." *Journal of Phonetics, 18,* 135–152.

Lisker, L. (1978). In qualified defence of VOT. *Language and Speech, 21,* 375–383.

Lisker, L., & Abramson, A. S. (1964). Cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20,* 384–422.

Lisker, L., & Abramson, A. S. (1970). The voicing dimension: Some experiments in comparative phonetics. In *Proceedings of the sixth international Congress of Phonetic Sciences* (pp. 563–567). Prague: Academia.

Lively, S. E., Pisoni, D. B., & Logan, J. S. (1992). Some effects of training Japanese listeners to identify English /r/ and /l/. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, production, and linguistic structure* (pp. 175–196). Tokyo: Ohmsha.

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,* 743–753.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify /r/ and /l/: A first report. *Journal of the Acoustical Society of America, 89,* 874–886.

McClaskey, C. L., Pisoni, D. B., & Carrell, T. D. (1983). Transfer of training of a new linguistic contrast in voicing. *Perception & Psychophysics, 34,* 323–330.

McClelland, J. L. (2001). Failures to learn and their remediation: A Hebbian account. In J. L. McClelland & R. S. Siegler (Eds.), *Mechanisms of cognitive development* (pp. 97–121). Mahwah, NJ: Erlbaum.

Nittrouer, S., & Miller, M. E. (1997). Predicting developmental shifts in perceptual weighting schemes. *Journal of the Acoustical Society of America, 101,* 2253–2266.

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

Nusbaum, H. C., & Goodman, J. (1994). Learning to hear speech as spoken language. In H. C. Nusbaum & J. Goodman (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 299–338). Cambridge, MA: MIT Press.

Nusbaum, H. C., & Lee, L. (1992). Learning to hear phonetic information. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, production, and linguistic structure* (pp. 265–273). Tokyo: Ohmsha.

Nusbaum, H. C., & Magnuson, J. (1997). Talker normalization: Phonetic constancy as a cognitive process. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 109–132). San Diego, CA: Academic Press.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In Y. Tohkura, Y. Sagisaka, & E. Vatikiotis-Bateson (Eds.), *Speech perception, production, and linguistic structure* (pp. 113–134). Tokyo: Ohmsha.

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. 113–158). San Diego, CA: Academic Press.

Pepperberg, I. M., & Neapolitan, D. M. (1988). Second language acquisition: A framework for studying the importance of input and interaction in exceptional song acquisition. *Ethology, 77,* 150–168.

Pickett, J. M. (1980). *The sounds of speech communication*. Boston: Allyn & Bacon.

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 109–132). San Diego, CA: Academic Press.

Pisoni, D. B., Aslin, R. N., Perey, A. J., & Hennessy, B. L. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception and Performance, 8,* 297–314.

Pisoni, D. B., Lively, J. S., & Logan, S. E. (1994). Perceptual learning of nonnative speech contrasts: Implications for theories of speech perception. In H. C. Nusbaum & J. Goodman (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 121–166). Cambridge, MA: MIT Press.

Polka, L. (1991). Cross-language speech perception in adults: Phonemic, phonetic, and acoustic contributions. *Journal of the Acoustical Society of America, 89,* 2961–2977.

Polka, L. (1992). Characterizing the influence of native language experience on adult speech perception. *Perception & Psychophysics, 52,* 37–52.

Rabiner, L. R., & Schafer, R. W. (1978). *Digital processing of speech signals.* Englewood Cliffs, NJ: Prentice Hall.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981, May 22). Speech perception without traditional speech cues. *Science, 212,* 947–950.

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

SAS Institute, Inc. (1997). *SAS/STAT software: Changes and enhancements through release 6.12.* Cary, NC: Author.

Schmidt, A. M. (1996). Cross-language identification of consonants: Part 1. Korean perception of English. *Journal of the Acoustical Society of America, 99,* 3201–3211.

Schroeder, L. D., Sjoquist, D. L., & Stephan, P. E. (1986). *Understanding regression analysis: An introductory guide* (Sage University Paper Series on Quantitative Applications in the Social Sciences, No. 07-057). London: Sage.

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

Sebeok, T., & Umiker-Sebeok, D. J. (1976). *Speech surrogates: Drum and whistle systems*. The Hague, the Netherlands: Mouton.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic

model relating generalization to distance in psychological space. *Psychometrika, 22,* 325–345.

Shepard, R. N. (1974). Representation of structure in similarity data: Problems and prospects. *Psychometrika, 39,* 373–421.

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

Smith, L. B., & Kemler, D. G. (1977). Developmental trends in free classification: Evidence for a new conceptualization of perceptual development. *Journal of Experimental Child Psychology, 24,* 279–298.

Steinschneider, M., Arezzo, J., & Vaughan, H. G., Jr. (1982). Speech evoked activity in the auditory radiations and cortex of the awake monkey. *Brain Research, 252,* 353–365.

Strange, W. (1995). Cross-language studies of speech perception: A historical review. In W. Strange (Ed.), *Speech perception and linguistic experience* (pp. 3–45). Baltimore: York Press.

Studdert-Kennedy, M., Liberman, A. M., Harris, K. S., & Cooper, F. S. (1970). Motor theory of speech perception: A reply to Lane's critical review. *Psychological Review, 77,* 234–249.

Sussman, H. M. (1988). The neurogenesis of phonology. In H. A. Whitaker (Ed.), *Phonological processes and brain mechanisms* (pp. 1–23). New York: Springer-Verlag.

Tees, R. C., & Werker, J. F. (1984). Perceptual flexibility: Maintenance or recovery of the ability to discriminate non-native speech sounds. *Canadian Journal of Psychology, 38,* 579–590.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327–352.

Tversky, A., & Gatti, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review, 89,* 123–154.

Walley, A. C., & Carrell, T. D. (1983). Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants. *Journal of the Acoustical Society of America, 73,* 1011–1022.

Werker, J. F., & Logan, J. S. (1985). Cross-language evidence for three factors in speech perception. *Perception & Psychophysics, 37,* 35–44.

Werker, J. F., & Tees, R. C. (1984). Phonemic and phonetic factors in adult cross-language speech perception. *Journal of the Acoustical Society of America, 75,* 1866–1878.

Yamada, R. A. (1995). Age and acquisition of second language speech sounds: Perception of American English /r/ and /l/ by native speakers of Japanese. In W. Strange (Ed.), *Speech perception and linguistic experience* (pp. 305–320). Baltimore: York Press.

Yamada, R. A., & Tohkura, Y. (1992). Perception of American English /r/ and /l/ by native speakers of Japanese. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, production, and linguistic structure* (pp. 155–174). Tokyo: Ohmsha.

Received February 9, 2000
Revision received March 28, 2001
Accepted July 17, 2001 ■