Journal of Experimental Psychology:
Human Perception and Performance
2009, Vol. 35, No. 6, 1913–1933

0096-1523/09/$12.00 DOI: 10.1037/a0015781

Supervised and Unsupervised Learning of Multidimensional Acoustic Categories

Martijn Goudbeek
University of Geneva and Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

Daniel Swingley
University of Pennsylvania

Roel Smits
Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

Learning to recognize the contrasts of a language-specific phonemic repertoire can be viewed as forming categories in a multidimensional psychophysical space. Research on the learning of distributionally defined visual categories has shown that categories defined over 1 dimension are easy to learn and that learning multidimensional categories is more difficult but tractable under specific task conditions. In 2 experiments, adult participants learned either a unidimensional or a multidimensional category distinction with or without supervision (feedback) during learning. The unidimensional distinctions were readily learned and supervision proved beneficial, especially in maintaining category learning beyond the learning phase. Learning the multidimensional category distinction proved to be much more difficult and supervision was not nearly as beneficial as with unidimensionally defined categories. Maintaining a learned multidimensional category distinction was only possible when the distributional information that identified the categories remained present throughout the testing phase. We conclude that listeners are sensitive to both trial-by-trial feedback and the distributional information in the stimuli. Even given limited exposure, listeners learned to use 2 relevant dimensions, albeit with considerable difficulty.

Keywords: auditory categories, supervised learning, unsupervised learning, nonspeech

Infants acquiring a first language and learners of a second language must learn to categorize the sounds of the language's phonetic system. To succeed, the learner must use phonetic information in the speech signal to determine how many categories there are and to categorize additional tokens of sounds as they are heard. Despite a consensus that this process should be conceptualized as a distributional learning problem (e.g., Guenther & Gjaja, 1996; P. K. Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Werker et al., 2007), little is known about the mechanisms by which category learning proceeds, or about what constraints on category learning are present (McCandliss, Fiez, Protopapas, Conway, & McClelland, 2002). The experiments presented here are first steps in a larger attempt to lay out general principles of auditory category learning, with particular reference to problems posed by phonetic categories (Francis & Nusbaum, 2002; Francis, Nusbaum, & Fenn, 2007; Holt & Lotto, 2006; McCandliss et al., 2002).

Our approach is similar to that taken in studies of visual category learning (Ashby & Maddox, 1993; Nosofsky, 1990), in which perceptual categories are defined as existing in a psychophysical space with continuous dimensions. We assume that when listeners hear a sound, this sound is evaluated on a number of dimensions and mapped onto a point in a multidimensional space. Repeated exposure to sounds originating from distributionally distinct categories leads to the formation of "clouds" of points. If, after a period of exposure, distinct clouds emerge, listeners can start to associate each cloud with a different category.

Most research on the learning of categories defined as clusters in perceptual space has investigated simple visual dimensions: the length and orientation of line segments, the slope of a line bisecting a circle and the size of the circle, the horizontal and vertical position of dots relative to a midline, and so forth. Here, we focus on the learning of similarly constructed auditory categories that are defined over simple auditory dimensions. Determining whether similar processes underlie category learning in different sensory modalities is itself of interest (e.g., Maddox, Ing, & Lauritzen, 2006). In addition, it is hoped that a better understanding of auditory category formation in tightly controlled experimental situations will inform theories of speech perception and language acquisition.

Martijn Goudbeek, University of Geneva, Geneva Emotion Research Group, Geneva, Switzerland, and Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands; Daniel Swingley, Department of Psychology, University of Pennsylvania; Roel Smits, Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands.

This research was carried out with the support of a Max Planck Society doctoral scholarship rewarded to Martijn Goudbeek as well as funding from the Swiss National Science Foundation (FNRS 101411–100367). Daniel Swingley was supported by National Institutes of Health Grant R01–HD049681 and by National Science Foundation Grant HSD–0433567 to D. Dahan and Daniel Swingley. The research further received support from the NWO–SPINOZA project "Native and Non-Native Listening" rewarded to Anne Cutler, who we thank for comments on the first version of the manuscript. Finally, we thank three reviewers for their helpful suggestions on the manuscript.

Correspondence concerning this article should be addressed to Martijn Goudbeek, Communication and Cognition, Faculty of Arts, Tilburg University, P.O. Box 90153, NL–5000 LE Tilburg, the Netherlands. E-mail: m.b.goudbeek@uvt.nl

We assume that recognition of the statistical patterns in the emerging clouds of points in multidimensional space is equivalent to category acquisition. The human capacity for resolving the categories of spoken language provides a particularly interesting example of perceptual learning because the acquisition of language-specific categories begins in infancy (Aslin, Jusczyk, & Pisoni, 1998; Jusczyk, 1997), and because this learning is necessarily unsupervised in nature. This last observation motivates the manipulation of the presence or absence of supervision (trial-by-trial feedback) in our experiments.

The distinction between supervised and unsupervised category learning has been explored extensively in adults. Human adults have proven adept at acquiring perceptual categories when given regular and immediate feedback about the validity of their judgments (Ashby & Alfonso-Reese, 1995; Ashby, Maddox, & Bohil, 2002; Francis, Baldwin, & Nusbaum, 2000; Gureckis & Love, 2003), but such feedback is not always required (Fiser & Aslin, 2001; Fried & Holyoak, 1984; Wade & Holt, 2005) and is seldom provided by everyday experience. When confronted with complex multidimensionally varying stimuli, learners must rely on the distributional structure of the objects and events they perceive. In successful perceptual categorization, those things that occupy nearby regions of perceptual space come to be regarded as the same, and as distinct from things that occupy different regions of this space. If an observer can detect the correlated structure of category members, he or she has a basis for forming a category without external feedback.

Unsupervised category learning studies have revealed characteristic limits in observers' abilities. Ashby, Queller, and Berretty (1999) showed that participants initially opt for unidimensional solutions (ignoring every dimension of variation but one) but can be brought to entertain multidimensional solutions with the aid of supervision. Several other studies also have shown the preference for the use of one dimension (Love, 2002) or of category structures with minor prototype distortions (Homa & Cultice, 1984). As stated previously, most of the evidence supporting these generalizations derives from experiments testing simple visual categories in which the dimensions of variation are readily identifiable to participants. Artificial categories involving distributions of more complex stimulus patterns which dimensions of variation are less obvious have rarely been used in unsupervised learning experiments, and, as suggested previously, few studies have used these methods to test the learning of auditory categories (but see Holt & Lotto, 2006; McCandliss et al., 2002; McClelland, Fiez, & McCandliss, 2002).

The literature on visual category formation suggests that in all likelihood, speech sound categories should be extremely difficult to learn. Not only do speech stimuli vary on many relevant dimensions, there is also considerable overlap between categories and variability within categories (e.g., Hillenbrand, Getty, Clark, & Wheeler, 1995; Peterson & Barney, 1952). Yet it is now well-known that infants are well on their way to learning the phonetic categories of their native language within the first year of life. Numerous experiments demonstrate the ability of infants to discriminate a broad range of speech sound contrasts early in development. Over the course of the first year infants start to conflate similar sounds if those sounds are not phonologically contrastive in the infant's native language (see, e.g., Aslin, Pisoni, & Jusczyk, 1983; Jusczyk, 1997, for reviews). Several studies have found decrements in nonnative consonant discrimination by the age of 12 months (e.g., Werker & Tees, 1984) and analogous decrements in nonnative vowel perception even earlier (P. K. Kuhl et al., 1992; Polka & Werker, 1994). These changes in discrimination ability are seen as adaptive for native language understanding because the failure to discriminate nonnative speech contrasts is taken to imply an improved understanding of the available speech categories in the native language (see P. K. Kuhl et al, 2006).

Thus, the improved recognition of speech categories of the native language may explain the loss of the infant's ability to discriminate nonnative phonemes, possibly because of changes in infants' attention to different phonetic cues. Once two nonnative sounds have become part of the same native category, it becomes more difficult to differentiate them from each other and their category comembers (Best, 1995). Within-category discrimination is more difficult than between-category discrimination because within-category sounds are heard as more similar to each other than between-category sounds (Cameron Marean, Werner, & Kuhl, 1992; P. K. Kuhl, 1985). Given that infants show evidence of perceptual knowledge of their native language before they can articulate any words (indeed, before many infants begin to babble), corrective feedback cannot be responsible for this learning. Retention of linguistically relevant phonetic contrasts based on semantically contrasting minimal pairs (words phonologically matching in all but one feature or segment) is also excluded for infants because infants' lexical knowledge is almost certainly too meager for language-specific phonological tuning to be driven by semantic contrast in phonologically similar words (Swingley, 2003). As a result, it is generally assumed that infants acquire their knowledge about phonetic categories via an unsupervised bottom-up distributional analysis of the speech they hear (e.g., Pierrehumbert, 2003).

A demonstration of such learning in a laboratory setting was provided in a study of 6- and 8-month-old infants by Maye, Werker, and Gerken (2002). In their study, two groups of infants were exposed to stimuli varying in formant trajectories, with prevoicing as a secondary cue on one end of the continuum. This led to a continuum extending from "da" to unaspirated "ta," a distinction not made in English. One group listened to stimuli in which the trajectories followed a unimodal distribution (most sounds were from the middle of the continuum) whereas the other group was presented with stimuli following a bimodal distribution (most sounds were from near the edges). Following this familiarization, infants were given the opportunity to listen to alternating stimulus sets (both of the endpoint stimuli) or nonalternating sets (the same stimulus repeated). Only the infants in the bimodal familiarization group evidenced a preference for nonalternating over alternating stimuli at test, revealing discrimination; infants in the monomodal group showed no such preference. Maye and Gerken (2000, 2001) found a similar sensitivity to distributional characteristics for adults with similar stimuli. However, the generality of this extremely rapid distributional learning is not clear at present (Peperkamp, Pettinato, & Dupoux, 2003; Pierrehumbert, 2003; Tyler & Johnson, 2006).

In the present contribution, we describe experiments in which adult listeners were tested on their ability to learn auditory categories. The categories comprised novel sounds with speech-like properties, to simulate processes of phonetic category learning

while minimizing effects of native-language phonological knowledge.

Our use of artificial categories exemplified by sampling from a distribution of variants of category prototypes ultimately descends from the pioneering studies of Attneave (1957) and Posner and Keele (1968), who laid out a range of hypotheses that are still of empirical interest. Among these are whether categories are abstracted as prototypes or stored as sets of experienced exemplars (or something in between), and when verbal descriptions of categories guide learners' decisions (see, e.g., Goldstone & Kersten, 2003). Here, we focused on two issues: first, how well listeners can learn two similar, distributionally defined auditory categories given limited supervised or unsupervised exposure; and second, how this learning is influenced by whether the category structures demand attention to one versus two dimensions of variation.

To generate our experimental stimuli, we specified a psychophysical space spanned by two acoustical dimensions known to be relevant in vowel perception, namely frequency and duration. Categories were defined as two-dimensional probability density functions in this space. Exemplars generated from these functions formed "clouds" in perceptual space. The statistical properties of the probability density functions (their means and covariance matrices) governed the relevance of each dimension for making category judgments (see Figure 1). For example, exposure to the structure in the top left cell in Figure 1 should encourage participants to categorize using only Dimension 1, and exposure to the structure in the bottom left cell should encourage participants to use only Dimension 2. In these unidimensional situations, the dimension that does not differentiate the categories is irrelevant to category assignment, although it contributes just as much to the variance of the probability density functions.

Exposure to the structures in the right-hand column should encourage the use of both dimensions in categorizing because the use of only one dimension would lead to many incorrect categorizations (Goudbeek, Swingley, & Kluender, 2007). Experiments in visual category learning have shown that participants initially prefer a unidimensional solution (Feldman, 2000) and only with the help of feedback start using a two dimensional strategy (Ashby, Alfonso-Reese, Turken, & Waldron, 1998). Ashby et al. (1998) distinguished between verbal- and procedural-based category learning. In their model, the verbal system has initial priority, and tries to categorize using a relatively simple (unidimensional) rule (e.g., long sounds in category A, short sounds in category B). Rules that are more complex and more difficult to verbalize like "all long and high frequency sounds go into category A" only enter the verbal system after the unidimensional rules have failed. The other category learning system in their model is an implicit or procedural learning system (Ashby & Waldron, 1999) that is based on the learning of actual skills or procedures (in this case, for categorization). This system does not have such a preference for unidimensional solutions but learns more slowly.

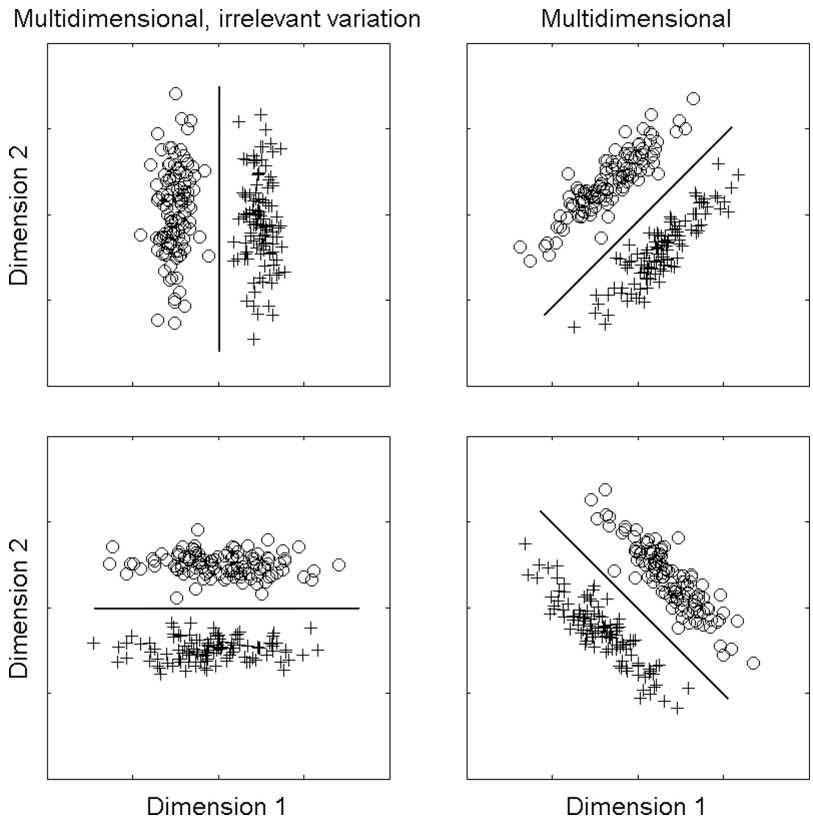

Figure 1. Four possible category structures in a two-dimensional perceptual space. Lines represent the optimal solution to the categorization problem.

The notion that learning categories defined over multiple dimensions could be more difficult than learning unidimensional categories may seem counterintuitive. Indeed, category learning is sometimes facilitated by the presence of multiple dimensions of variation. When multiple cues are available to aid in the identification of a category member, or when nominally distinct dimensions' values are interpreted holistically, redundancy gain may be observed (e.g., Egeth & Mordkoff, 1991; Garner, 1974; Pomerantz & Lockhead, 1991). In addition, the presence of correlated attributes among some members of a set of objects can lead observers to form a category that includes those members and excludes the rest—an effect that has been demonstrated even in 10-month-old infants (Younger, 1985). However, these advantages of correlations among stimuli depend on redundancy. Note that in the "diagonal" categories in the right-hand column of Figure 1, the value of only one dimension is not a reliable predictor of category membership; good performance requires use of both dimensions. Relative to unidimensional "filtering" tasks (left-hand column), any advantage due to correlations among the dimensions may be outweighed by the fact that listeners must attend to two dimensions rather than one. Thus, the multidimensional-categorization task (sometimes referred to as a *condensation* task) is more difficult than analogous unidimensional tasks (Gottwald & Garner, 1972; Posner & Keele, 1970).

Distinguishing diagonal and nondiagonal category distributions presupposes the psychological reality of the axes and a particular interpretation of the axes' orientation. This notion has been studied in attempts to understand the separability or integrality of pairs of dimensions. Broadly speaking, two separable dimensions can be attended to exclusively without mutual interference, although integral dimensions cannot (Garner, 1974). This leads to the prediction that if two category sets defined along separable dimensions are rotated in stimulus space (converting the left column of Figure 1 to the right column), categorization should become substantially more difficult because observers are deprived of the effective strategy of ignoring the irrelevant dimension (or, conversely because any tendency to rely on a single dimension leads to many errors). This prediction has been upheld in a number of studies, although the situation is complicated by the fact that classification of dimension pairs as separable or integral is not always maintained consistently over tasks (more thorough discussion of these issues may be found in Grau & Kemler Nelson, 1988; Kemler Nelson, 1993; Melara & Marks, 1990; Shepard, 1991). To anticipate our results, the present experiments reveal a large axis rotation effect, revealing that the speech-like dimensions under study are "psychologically real" in Grau and Kemler Nelson's sense.

In our experiments adult listeners were exposed to categories of nonspeech sounds. These were inharmonic tone complexes filtered by a single resonance. The two dimensions of variation were the frequency of the spectral peak at which the sound complex was filtered (formant frequency) and the duration of the stimulus (duration). These dimensions are important in the perception of vowel sounds (e.g., Ainsworth, 1972; Peterson & Barney, 1952).

Although in principle models of language acquisition might best be developed using novel speech categories (such as phonetic categories not present in the language of the participants), it is well-known that users of a given language tend to interpret sounds from nonnative languages in terms of the perceptual categories of their native language (Best, McRoberts, & Sithole, 1988; Best & Strange, 1992; Flege, 1995; Polivanov, 1931) especially after being trained to identify these stimuli (Francis et al., 2007). This complicates efforts to model category acquisition in naïve listeners, and motivated our choice to use nonspeech sounds as stimuli. However, because these dimensions (or closely related ones) are necessary for speech interpretation, there is no reason to expect that success in the task would require the development of genuinely novel features or stimulus dimensions (see Francis & Nusbaum, 2002, for discussion and evidence bearing on this point for speech sounds; Schyns, Goldstone, and Thibaut, 1998, regarding feature creation more general). For example, given that the native language of the participants was Dutch (Booij, 1995), all participants were fully accustomed to distinguishing the vowels in words like *maan* ("moon"), *man* ("man"), and *men* ("people"). The first two words' vowels may differ primarily in their duration (Nooteboom & Doodeman, 1980), although the last two words' vowels differ in their formant frequencies. Thus, although the inharmonic tone complexes did not sound like spoken words, the dimensions of variation themselves were not new.

Listeners' exposure to the category structures was given through experience with category exemplars, in a forced-choice decision task with feedback on each trial in Experiment 1, and without trial-by-trial feedback in Experiment 2. The supervised learning procedure in Experiment 1 thus was comparable to the typical procedure used in visual category learning studies and in speech-contrast training studies (e.g., Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1997; Greenspan, Nusbaum, & Pisoni, 1988; Lively, Logan, & Pisoni, 1993). The unsupervised learning procedure in Experiment 2 was more comparable to the situation of infants learning their first language. Learning of multidimensionally varying categories with relevant variation in one dimension was tested in Conditions 1 and 2 of each experiment, whereas learning of multidimensional categories with relevant variation in two dimensions was tested in Condition 3.

All experiments used the same basic procedure, with a learning phase and a maintenance phase. In the learning phase, listeners were presented with stimuli drawn from two probability density functions. They were faced with the problem of partitioning the psychophysical space by using a criterion based on one or more dimensions. Listeners' use of a unidimensional criterion would be reflected in their assignment of all stimuli below a criterion value on that dimension to one category, and all stimuli above it to another (Ashby & Maddox, 1990). The use of a multidimensional criterion would be reflected by listeners' allowing dimensions to trade off: For example, a low value on one dimension might be compensated by a low value on the other (or a high value on the other, depending on the orientation of the category's "diagonal" in perceptual space). This compensation entails interpretation of one dimension relative to the value of the other in assigning category membership—a process that is a hallmark of speech perception (e.g., Repp, 1982). In Conditions 1 and 2 the categorization problems could be solved completely (no miscategorized stimuli) by using one dimension, although the categorization problem of Condition 3 (and Experiment 1B) required the use of both dimensions for good categorization.

After the learning phase, participants entered a maintenance phase intended to characterize their division of psychophysical space. The stimuli of all maintenance phases except those of

Experiment 1B were drawn from an equidistantly spaced grid that was intended to "scan" the participants' psychophysical space in a neutral way, without continued distributional information (see the lower right panel of Figure 2). This change in stimulus properties permitted more accurate assessment of listeners' use of each dimension of variation, and also allowed evaluation of whether participants would maintain their category identification criteria once the distributional cues to category membership were no longer supported in the input. In Experiment 1B, we compared maintenance performance on this grid with maintenance of the learned category identification criteria on the same stimuli as in the learning phase. In none of the maintenance phases did the listeners receive trial-by-trial feedback.

Experiment 1: Supervised Learning

Method

Participants. Thirty-six participants (12 in each condition), all students from the University of Nijmegen, were drawn from the Max Planck Institute (MPI) participant pool and participated in return for a small payment. None of the participants reported any history of hearing problems.

Stimuli. The stimuli were inharmonic sound complexes, 112 in each category. All stimuli were created by modifying a base signal. This base signal was an inharmonic sound complex made by adding several sinusoids with exponentially spaced frequencies. The base signal was defined by the following formula:

$$B(t) = A \sum_{n=0}^{N-1} \sin(2\pi f_0 F^n t) \quad (1)$$

where A represents the amplitude of the signal, f_0 is the frequency of the lowest sinusoid (500 Hz), t is time in seconds, and F is the frequency ratio between two successive sinusoids (1.15). Thus, the frequencies of the base signal were not spaced linearly, as they are in harmonic (e.g., speech) sounds. Finally, N is the total number of sinusoids that were added together; this was set to 17.

After the base signal was constructed, it was filtered with a single resonance peak, implemented as a second order infinite impulse response (IIR) filter. The filter's bandwidth was 0.2 times that of its resonance frequency. Each sound was truncated at the desired duration, applying linear onset and offset ramps of 5 ms to avoid the perception of clicks.

In all experiments, the stimuli varied in two dimensions: the frequency of the spectral peak at which the sound complex was filtered (our nonspeech analogue of formant frequency) and the duration of the sound. To ensure that both dimensions would be equally salient and discriminable, they were converted to psychopysical scales and normalized using their respective just noticeable differences (JND). The psychophysical scale commonly accepted for the perception of frequency is the equivalent rectangular bandwidth scale (Glasberg & Moore, 1990). With this scale, physical frequency f expressed in Hertz is transformed to "psychological frequency" e expressed in equivalent rectangular bandwidth (ERB) units as follows:

$$e = 21.4^{10}\log(0.00437 \times f + 1) \quad (2)$$

Psychological duration DUR (measured in psychophysical units coined DUR) is converted from stimulus duration t (expressed in seconds) according to the following transformation:

$$D = 10\log(t) \quad (3)$$

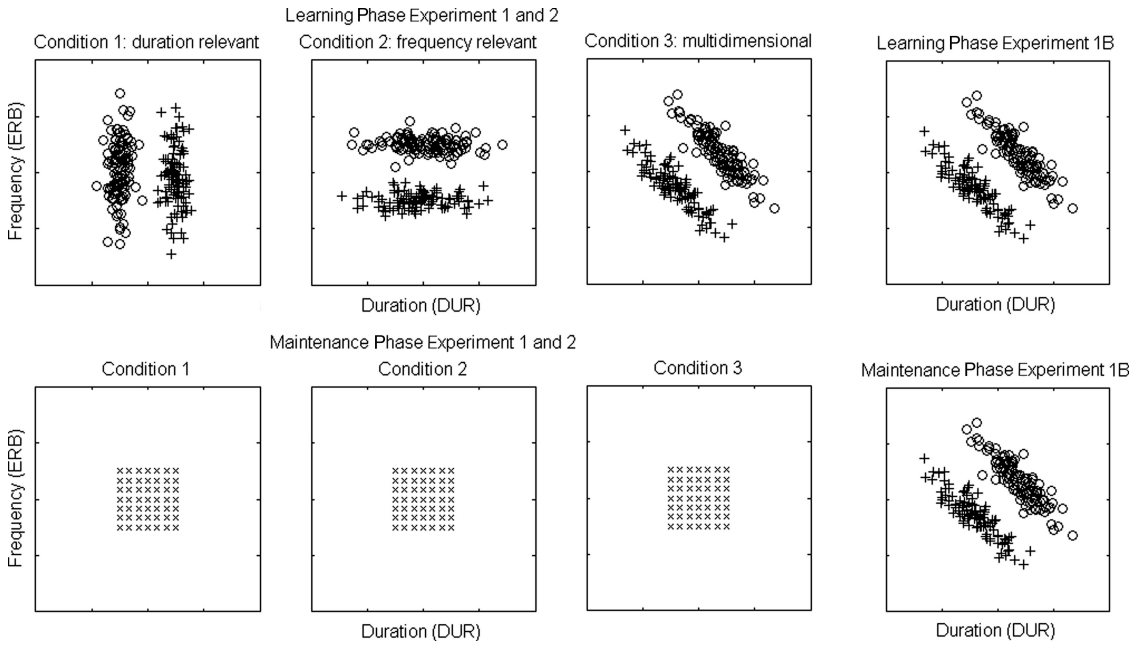

Figure 2. Learning (upper panels) and maintenance (lower panels) conditions of Experiments 1 and 2 and the learning and maintenance conditions of Experiment 1B (rightmost panels). ERB = equivalent rectangular bandwidth.

This transformation was proposed by Smits, Sereno, and Jongman (2006) based on data published by Abel (1972). The relevant JND in this frequency region for formant frequency is 0.12 ERB (Kewley-Port & Watson, 1994). For duration, experiments by Smits et al. and subsequent piloting with multidimensional stimuli varying in duration and frequency indicated that a JND of 0.25 DUR resulted in a discriminability comparable to 0.12 ERB. We used these values to equalize the range of variation between the stimulus dimensions, so that the difference between the category means in the training distributions and between the highest and the lowest stimulus value in the grid used in the Maintenance Phase was 20 JNDs for both frequency and duration.

Our stimuli are constructed in the same way as those used by Smits et al. (2006). The participants in their experiment, who were drawn from the same MPI-participant pool, typically described the stimuli as sounding like computer sounds, organs, or horns (Smits et al., 2006). Figure 3 contains spectrograms of four stimuli used in the experiment. The spectrograms in Figure 3 depict stimuli of short and long duration and of high and low frequency, spanning the whole range of stimuli used in our experiment. As the spectrograms imply, the stimuli varied in dimensions relevant for speech sound identification, but would not be confused for or interpreted as actual speech sounds.

Solving the categorization problem in Conditions 1 and 2 required the use of only one dimension, whereas solving the problem in Condition 3 required the use of both dimensions. In Condition 1, the stimuli manifested relevant variation in duration and irrelevant variation in formant frequency. In Condition 2, the stimuli manifested relevant variation in formant frequency and irrelevant variation in duration. In Condition 3, the stimuli manifested relevant variation in both dimensions (see the first three upper panels of Figure 2). To ensure a large enough incentive for participants to actually use both dimensions in Condition 3 (Goudbeek et al., 2007), we chose the mean and covariance matrices of the two distributions such that using a unidimensional solution to the categorization problem resulted in a much lower optimal percentage of correctly categorized stimuli (70%) than using the optimal two-dimensional solution (100%). Table 1 shows the perceptual and physical characteristics of the distributions of the learning stimuli of each condition.

The maintenance stimuli were the same for all conditions, with items taken from an equidistantly spaced grid (see the lower left panels of Figure 2 and Table 2).

Procedure. Participants were seated in a soundproof booth in front of a computer screen and a two-button response box. In the learning phase, they listened to 448 stimuli (two categories times 112 stimuli per category times two presentations) through Sennheiser closed-ear headphones (Sennheiser, Almere, the Netherlands). The stimuli from the two categories were presented in a random order in two blocks separated by a brief rest period. All 112 stimuli from each category were presented once in each block.

The listeners' task was to assign each stimulus to group A or B, using the button box. When their categorization was correct, the monitor displayed (the Dutch equivalent of) "right" in green letters for 700 ms; when the categorization was incorrect, the monitor displayed (the Dutch equivalent of) "wrong" in red letters for 700 ms immediately following the response. After the visual feedback disappeared, a 200-ms blank screen preceded the next stimulus.

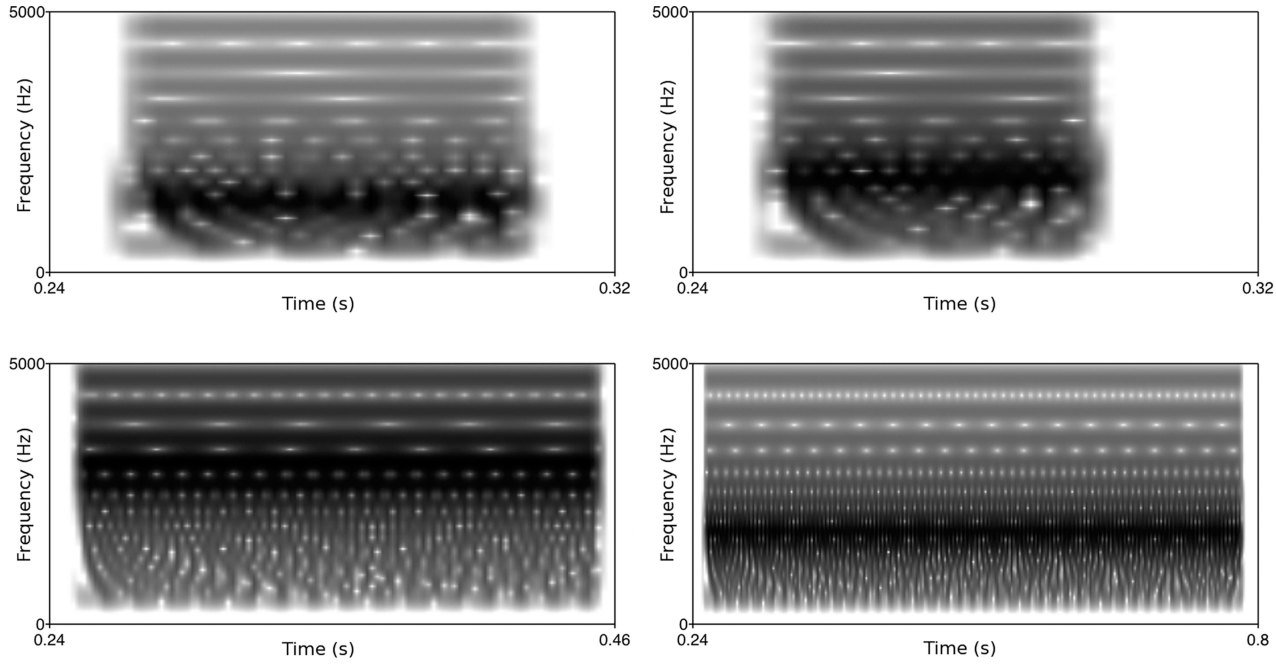

Figure 3. Spectrograms of four stimuli used in the experiment. Note the different time scales due to differences in stimulus duration. Listeners reported stimuli as being similar to speech, but definitely nonspeech (Smits, Sereno, & Jongman, 2006).

Table 1
Distributional Characteristics of the Learning Stimuli With Relevant Variation in One Dimension (Condition 1 and 2) or Relevant Variation in Two Dimensions (Condition 3)

	Category A			Category B		
Variable	*M*	σ	ρ	*M*	σ	ρ
Condition 1 (duration relevant)	47.7 DUR (117 ms) 18.80 ERB (1501 Hz)	0.65 DUR (1.07 ms) 1.88 ERB (51.3 Hz)	−.05	52.53 DUR (205.0 ms) 18.90 ERB (1520 Hz)	0.65 DUR (1.07 ms) 1.88 ERB (51.3 Hz)	−.10
Condition 2 (frequency relevant)	50.1 DUR (149.6 ms) 17.6 ERB (1295 Hz)	6.45 DUR (1.91 ms) 0.31 ERB (7.76 Hz)	.05	49.73 DUR (144.5 ms) 20.0 ERB (1737 Hz)	6.46 DUR (1.91 ms) 0.31 ERB (7.76 Hz)	.10
Condition 3 (multidimensional)	48.38 DUR (126.2 ms) 17.79 ERB (1322 Hz)	2.80 DUR (1.32 ms) 1.34 ERB (35.5 Hz)	−.98	51.66 DUR (175.2 ms) 19.70 ERB (1977 Hz)	2.82 DUR (1.33 ms) 1.33 ERB (35.2 Hz)	−.98

Note. DUR = psychophysical unit for perceived duration; ERB = equivalent rectangular bandwidth.

In the maintenance phase participants categorized sounds from the test continuum, as belonging to group A or B. There were 49 maintenance stimuli that were randomly ordered in four blocks, totaling 196 presentations. Once a participant had selected a category label on a trial, the monitor would display (the Dutch equivalent of) "next" for 700 ms and the next stimulus was played after a 200-ms delay. No feedback was given on maintenance trials.

Results and Discussion

The results were analyzed using percentage correct, d' and logistic regression. Both d' and percentage correct are familiar measures of performance. A disadvantage is that they are based on category membership and not on the coordinates of each individual stimulus in the duration/formant-frequency plane and consequently they yield less fine-grained information about participants' strategies. In addition, they cannot be applied to the data of the maintenance phase because correctness of a response does not apply straightforward in the region between the trained category exemplars. Logistic regression, on the other hand, is sensitive to the coordinates of the stimuli, and can be applied to the data of the maintenance phase (Agresti, 1990).

In regression analysis, linear and interaction terms can be entered into the analysis. For the present kind of analysis, the interpretation of an interaction term is often problematic, and is usually left out in studies of this type. Here, the results were analyzed both with and without the interaction term. Of the 144 analyses in Experiments 1 and 1B (12 participants × 4 analysis conditions × 3 experimental parts) only 12 had a significant interaction term. Furthermore, the fits of the models with interaction term hardly improved compared to those without. Based on these results we present here only the model without the interaction term.

Signal detection analysis (percentage correct and d′). The data of the learning phases were analyzed first using percentage correct and d'. To probe for learning, the first and second halves of the learning phase were analyzed separately. Listeners' performance was fairly good. The three upper rows of Table 3 show the percentages correct and d' values of the first and second part of the learning phase of Condition 1 (duration relevant), Condition 2 (frequency relevant) and Condition 3 (multidimensional learning). Recall that percentage correct and d' were only computed for the learning phase because it is there that right and wrong can be clearly assigned. In all conditions and both learning phases, percentages correct, and d's were significantly above chance (all $p <$ .05) in t tests with correction for multiple comparisons.

An analysis of variance (ANOVA) with part of the experiment (Learning Phase 1 vs. 2) as a within-subjects variable and condition (duration relevant vs. frequency relevant vs. multidimensional) as a between-subjects variable revealed significant improvements in performance from the first phase to the second, for the percentage correct measure, $F(1, 33) = 29.27$, $p < .05$, $\eta_p^2 = 0.47$; and the d' measure, $F(1, 33) = 33.29$, $p < .05$, $\eta_p^2 = 0.50$. Both analyses showed a significant difference between conditions, $F(1, 2) = 43.10$, $p < .05$, $\eta_p^2 = 0.63$; and $F(1, 2) = 28.36$, $p < .05$, $\eta_p^2 = 0.72$; for percentage correct and d', respectively. Post hoc multiple comparisons (Tukey's honestly significant difference [HSD]) showed no significant differences between the unidimensional conditions, although Condition 3 differed significantly from both Condition 1 and 2, indicating the advantage of unidimensional learning over multidimensional learning. Follow-up analyses conducted for each condition separately revealed significant differences between the first and second parts of the experiment for both percentage correct, $F_{min}(1,11) = 6.23$, $p < .05$, $\eta_p^2 = 0.36$; and d', $F_{min}(1,11) = 8.78$, $p < .05$, $\eta_p^2 = 0.44$; for all conditions. The signal detection measures thus indicated that learn-

Table 2
Distributional Characteristics of the Maintenance Phase (Equidistantly Spaced Grid)

Variable	*M*	Minimum	Maximum	Step size
Duration	50.1 DUR (150 ms)	47.6 DUR (117 ms)	52.6 DUR (193 ms)	0.84 DUR/step (12.7 ms/step)
Formant frequency	18.8 ERB (1499 Hz)	17.6 ERB (1288 Hz)	20.00 ERB (1739 Hz)	0.4 ERB/step (75.17 Hz/step)

Note. DUR = psychophysical unit for perceived duration; ERB = equivalent rectangular bandwidth.

Table 3
Signal Detection Results (Mean Percentage Correct and d′) With Their Standard Deviations for Experiments 1 and 1B

	Learning Phase 1				Learning Phase 2			
Variable	% correct	σ	*d′*	σ	% correct	σ	*d′*	σ
Experiment 1, Condition 1	0.81	0.04	1.39	0.21	0.93	0.02	2.59	0.27
Experiment 1, Condition 2	0.80	0.03	1.32	0.17	0.89	0.03	2.07	0.25
Experiment 1, Condition 3	0.59	0.01	0.33	0.05	0.63	0.01	0.50	0.05
Experiment 1B	0.58	0.02	0.28	0.08	0.62	0.03	0.45	0.11

ing a multidimensional distinction was feasible, but significantly more difficult than learning a unidimensional one.

Logistic regression. Logistic regression yields two β weights, similar to the weights in a linear regression, that reflect the influence of the independent variables (here, the perceptual dimensions) on the dependent variable (the listener's choice). A β weight of large magnitude indicates a strong influence of the associated dimensions on the dependent variable. The β weights were calculated separately for each participant. Comparing the effects of β weights for unidimensional (Condition 1 and 2) and multidimensional (Condition 3) learning problems is problematic because of conflicting predictions for successful unidimensional versus multidimensional performance. For this reason, Conditions 1 and 2 are analyzed separately from Condition 3.

Table 4 and Figure 4 display the mean β weights for the relevant and irrelevant dimension of Condition 1 and 2 for the first half of the learning phase (Learning Phase 1), the second half of the learning phase (Learning Phase 2) and the Maintenance Phase.

In addition to β weights, the logistic regression gives significance levels of the hypothesis that each β weight differs from zero. If a β weight did not differ significantly from zero at the $p = .05$ level, we concluded that participants did not make use of that dimension. The columns of Table 4 labeled "Unidimensional" and "Multidimensional" show how many participants used either one or both dimensions significantly. Numbers of participants who did not use any dimension significantly are not shown (note that the number of participants in each group was always 12).

Table 4 and Figure 4 confirm that in both conditions participants learned to use the relevant dimension. Both the mean β weights and the number of participants using that dimension were higher than those of the irrelevant dimension. This also shows that participants did not make systematic use of the irrelevant dimension of variation in making their judgments, as the values of the irrelevant dimensions remained close to zero throughout the experiment. The higher mean β weights and number of listeners using the relevant dimension in Condition 2 compared to Condition 1 suggest that formant frequency was an easier dimension to learn to attend to than duration. In the maintenance phase, when feedback was no longer given and the stimulus grid was used, listeners persisted in their use of the relevant dimensions. However, although formant frequency was easier to learn, it also appeared easier to unlearn, as was evidenced by the large drop in the average β weight for formant frequency in the maintenance phase.

To statistically test these effects, we carried out an ANOVA with part of the experiment (Learning Phase 1, Learning Phase 2, and Maintenance Phase) and dimension (relevant vs. irrelevant) as within-subjects variables, and condition (duration relevant vs. formant frequency relevant) as between-subjects variable and the β weights as dependent measures.

Because of a significant three-way interaction between dimension, part of the experiment and condition, the results were further analyzed for each condition separately.[1] For Condition 1 (duration relevant), the β weight for the relevant dimension was higher than that for the irrelevant dimension, $F(1, 11) = 61.06$, $p < .05$, $\eta_p^2 = 0.85$, which confirmed that listeners learned to attend to the relevant dimension. The significant main effect for part of the experiment, $F(2, 22) = 12.83$, $p < .05$, $\eta_p^2 = 0.54$, shows that participants improved over the course of the training. The interaction between part of the experiment and dimension, $F(2, 22) = 14.40$, $p < .05$, $\eta_p^2 = 0.57$, indicates that the learning effect depended on whether a dimension was relevant or irrelevant: The effect for part of the experiment was present for the relevant dimension, $F(2, 22) = 13.78$, $p < .05$, $\eta_p^2 = 0.56$, but not the irrelevant dimension, $F(2, 22) = 1.69$, *ns*, $\eta_p^2 = 0.13$.

In Condition 2, the same main effects and interactions as in Condition 1 were present. The β weight for the relevant dimension (frequency) was higher than that of the irrelevant dimension, $F(1, 11) = 175.04$, $p < .05$, $\eta_p^2 = 0.94$; and this advantage for the relevant dimension increased during the learning phase: part of experiment effect, $F(2, 22) = 15.61$, $p < .05$, $\eta_p^2 = 0.59$. The interaction between part of the experiment and dimension was also present; post hoc analysis showed a significant effect of part of the experiment for the relevant dimension, $F(2, 22) = 17.34$, $p < .05$, $\eta_p^2 = 0.61$, and a much smaller though significant effect for the irrelevant dimension, $F(2, 22) = 3.54$, $p < .05$, $\eta_p^2 = 0.24$. This difference between the conditions is caused by differences in their maintenance phases. In Condition 1, when duration was the relevant condition, its β weight remained high in the Maintenance Phase and the β weight for frequency remained small. In Condition 2 however, the β weight for frequency dropped in the Maintenance Phase and that of duration rose. Thus, even when they had previously correctly used formant frequency, listeners had a tendency to start using duration again when presented with an evenly spaced stimulus grid and without feedback.

The difference between learning to use and maintaining the use of duration and frequency was unexpected, particularly given our attempt to equalize the tested dimensions by scaling the variability

[1] The main effects of part of the experiment, $F(1, 22) = 13.85$, $p < .05$, $\eta_p^2 = 0.39$; dimension, $F(2, 44) = 187.98$, $p < .05$, $\eta_p^2 = 0.90$; but not that for condition, $F(1, 22) = 0.003$, *ns*, $\eta_p^2 = 0.00$; were significant as were all other interactions.

Table 4
Logistic Regression Results of Experiment 1 for Conditions 1 and 2

	Condition 1 (duration relevant)				Condition 2 (frequency relevant)			
Variable	μ (β)	σ (β)	Unidimensional	Multidimensional	μ (β)	σ (β)	Unidimentional	Multidimensional
Learning Phase 1								
Relevant	0.65	0.13	10	0	1.37	0.73	11	1
Irrelevant	0.05	0.04	0		0.02	0.03	0	
Learning Phase 2								
Relevant	1.50	0.27	11	0	2.28	1.11	11	1
Irrelevant	0.10	0.10	0		0.02	0.04	0	
Maintenance Phase								
Relevant	1.54	0.14	12	0	0.20	0.18	9	1
Irrelevant	0.10	0.06	0		0.07	0.06	0	

Note. Mean β weights are shown for both dimensions and the number of participants out of 12 using one (unidimensional) or both (multidimensional) dimensions significantly.

of the stimuli to empirically determined JNDs. Apparently, the similar JNDs obtained using same/different experiments varying one dimension in a two-dimensional formant-frequency × duration space did not guarantee equal categorization behavior. Smits et al. (2006) found a similar difference and hypothesized that it may be due to a difference in stimulus dimensions introduced by Stevens and Galanter (1957). Stevens and Galanter argued that dimensions like duration are prothetic dimensions, for which an increase in value means adding more of the same, whereas dimensions like formant frequency are metathetic dimensions, in which an increase does not necessarily mean more of the same. According to the model proposed by Smits et al., storing a category representation or comparing a stimulus with a stored category based on a prothetic dimension is noisier than storing a category representation or comparing a stimulus with a stored category based on a metathetic dimension and thus more difficult in the absence of feedback. This description is consistent with our (unidimensional) results.

Another possibility is that duration and frequency were differentially available to the participants in these stimuli. That is, to a first approximation the duration of a signal bounded by silence may be measured in a similar way regardless of the spectral characteristics of the signal; but extracting the peak frequency of these tone complexes may have been intrinsically more difficult, or may have profited less from participants' background experience in processing auditory signals. Although speech makes use of frequency peaks broadly similar to those tested here (and listeners are exquisitely sensitive to variations in these speech features), the present stimuli were not speech signals. If the participants' estimation of frequency was noisier than their estimation of duration, this could have led to their relative disregard for frequency in the Maintenance Phase (see, e.g., Zwicker & Fastl, 1990, pp. 265–271). We will return to this issue in Experiment 1B, in which the effect of the distributional information in the Maintenance Phase on the use of these dimensions will be investigated.

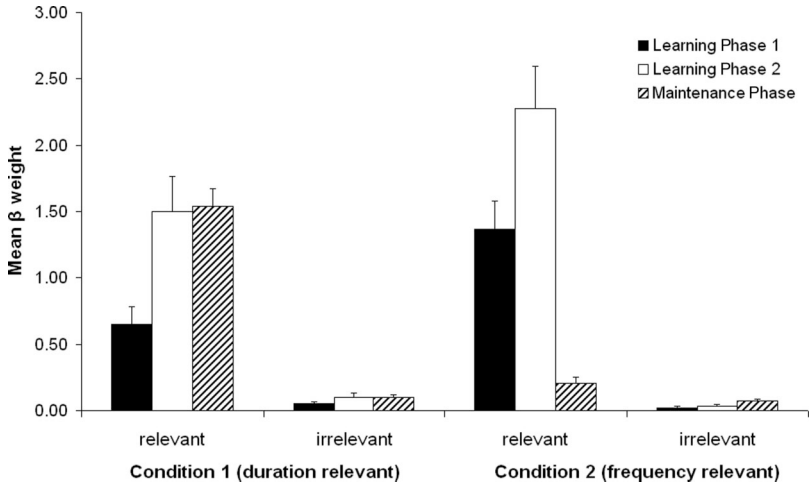

Figure 4. Mean β weights of Condition 1 and Condition 2 of Experiment 1 for the relevant and irrelevant dimensions for each part of the experiment. In Condition 1, duration was the relevant dimension of variation; in Condition 2, formant frequency was relevant. Vertical line segments indicate plus one standard error.

In summary, these data show that listeners can, relatively quickly, learn a unidimensional categorization in a two-dimensional space and generalize this learning to new exemplars, though this learning is not always robustly maintained.

Condition 3 addressed learning of multidimensional categories with two relevant dimensions of variation. Instead of what was effectively a unidimensional distinction in Condition 1 and 2, participants of Condition 3 had to learn a truly multidimensional distinction: both duration and formant frequency had to be used to obtain a high level of correct responding. Given that our interest is in whether individual participants used both dimensions (and not, say, half using one and half using the other), we present the results of Condition 3 as a set of scatterplots in which each point corresponds to one participant. The left-hand side of Figure 5 presents the β weights for duration (abscissa) and formant frequency (ordinate) for each listener in each part of the experiment. The data points are divided into four groups: listeners who used both dimensions significantly (identified by asterisks), listeners who used only formant frequency

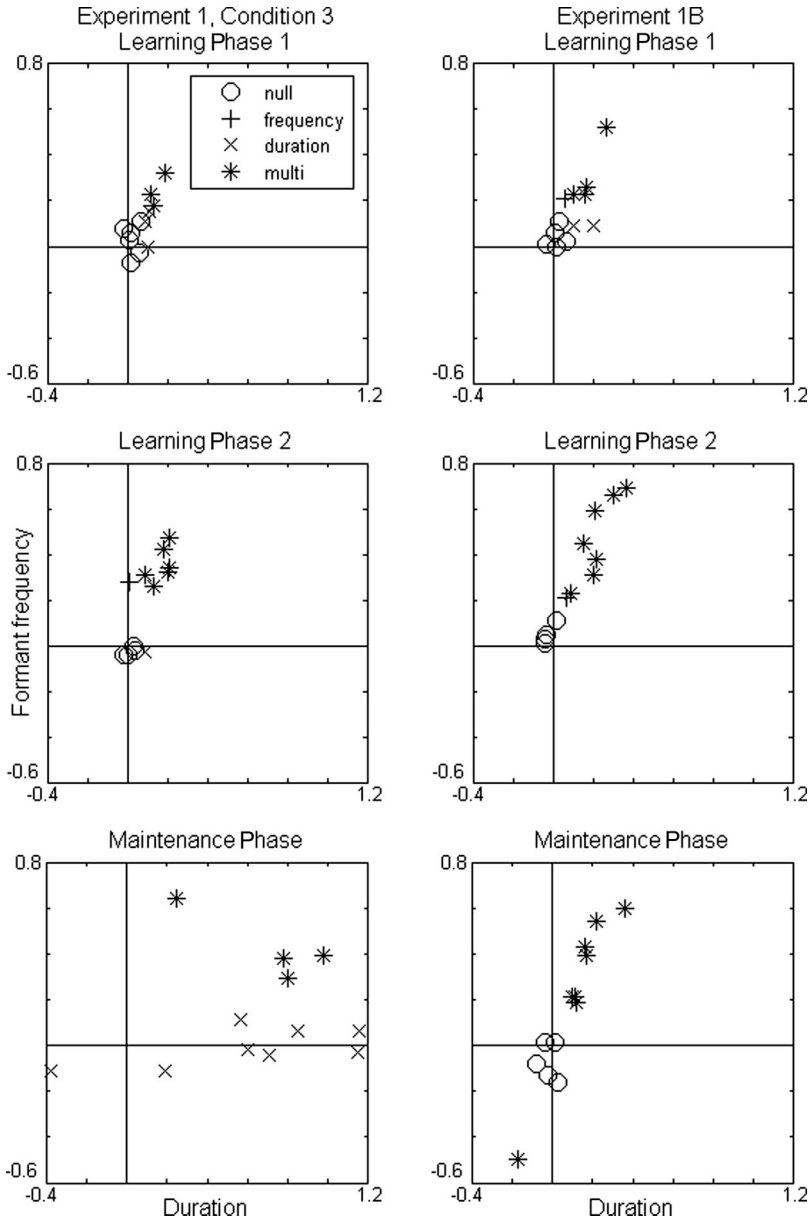

Figure 5. The listeners' β weights associated with the duration and formant-frequency dimensions for the two-dimensional categorization problems of Condition 3 of Experiment 1 (equidistantly spaced grid in the maintenance phase) and Experiment 1B (distributional information in the Maintenance Phase). Plotted in a two-dimensional duration-frequency plane. Asterisks indicate participants who used both dimensions significantly, Xs indicate listeners who used only duration significantly, plusses indicate listeners who used formant frequency significantly, and listeners marked by circles did not use any dimension significantly.

(plus signs), listeners who used only duration (Xs), and listeners who did not use any dimension significantly (circles). Optimal performance corresponds to a point in the upper right-hand corner of the square, at an angle of 45° (when both dimensions are given equal weight) and far away from the origin (reflecting high β weights and thus consistent behavior).

The two upper panels of the left-hand column of Figure 5 show performance in the first and second learning phase of Condition 3. Judging by the number the asterisks a number of listeners picked up on the information provided by the shapes of the categories' distributions and the feedback. Improvement in the second part is evident in the higher β values (i.e., asterisks closer to the upper-right corner). However, the third panel shows that listeners had trouble maintaining their learned categorization strategy (only four asterisks remain in the Maintenance Phase) and started using a unidimensional rule with duration as the relevant dimension (the Xs).

Most participants succeeded in using one or more dimensions above chance levels, whereas some failed to use any dimension significantly. For the purpose of comparing the performance of the successful participants across conditions and experiments, it would be desirable to have a measure of these participants' central tendency and variability. Note that simply computing the across-subjects average β weights for each of the dimensions would not be an effective way to characterize overall performance. For example, if half of these participants used duration exclusively, and the others formant frequency, the average β weights for duration and frequency might both exceed chance even though none of the individuals actually used both dimensions. These considerations suggest that a measure that integrates performance on both dimensions would be useful.

Here, we derive such a measure by computing the angle formed by the line connecting each participant's β weights to the origin, and also computing the length of this line. These computations were done by transforming the Cartesian coordinates of the β weights for duration and formant frequency into the polar coordinates Φ (the angle with the horizontal axis in radians) and A (the distance to the origin) by the following transformations:

$$A = \sqrt{(\beta_{dur}^2 + \beta_{freq}^2)} \tag{4}$$

$$\varphi = \arctan(\beta_{freq}/\beta_{dur}) \; if \; \beta_{dur} > 0 \tag{5a}$$

$$\varphi = \arctan(\beta_{freq}/\beta_{dur}) + \pi \; if \; \beta_{dur} \leq 0 \tag{5b}$$

$$\varphi: = \varphi - 2\pi \; if \; \varphi > \pi \tag{5c}$$

In our analysis, Φ ranges between π and –π radians. When Φ equals (1/2)π, listeners purely use formant frequency, when Φ equals 0, listeners use only duration, and when Φ is close to (1/4)π participants are in between those two angles and use duration as well as formant frequency. As can be seen from Figure 5, listeners who used both dimensions fall in the upper right-hand plane, somewhere between 0 and (1/2)π.

The other polar coordinate, A, ranges between zero and infinity. A large A indicates that a participant was internally consistent (though a large average A over participants need not reflect consistent weights of each dimension across participants); whereas a small A indicates that listeners' categorizations tend not to be internally consistent. In Figure 5, the listeners that categorized using both dimensions (indicated by the asterisks) are farther removed from the origin, while listeners that do not use any dimension significantly (the circles) are all very close to the origin. The left column of Table 5 lists the mean values of Φ for each phase of Condition 3 for all participants who in a given phase used one or more dimensions above chance levels.

The mean Φ of the first learning phase differed significantly from 0, $t(5) = 5.12$, $p < .05$ as well as from (1/2)π, $t(5) = -4.73$, $p < .05$. In the second learning phase, mean Φ was again significantly different from both 0, $t(7) = 4.96$, $p < .01$; and (1/2)π, $t(7) = -2.88$, $p < .05$. Mean Φ values exceeded (1/4)π (the value that would reflect an unbiased use of duration and formant frequency), indicating a somewhat stronger use of the frequency dimension than the duration dimension. As a group, participants used only duration in the Maintenance Phase of Condition 3. The mean Φ for participants using any dimension was not significantly different from 0, $t(11), = -0.243$, *ns;* but did differ significantly from (1/2)π, $t(11) = -5.850$, $p < .01$.[2]

An ANOVA with A as the dependent variable and part of the experiment as a within-subjects variable showed a significant effect of part of the experiment, $F(2, 10) = 5.863$, $p < .05$, $\eta_p^2 = 0.54$. Pairwise comparisons showed this effect to be due to a significant difference between the second[3] Learning Phase and the Maintenance Phase ($p < .05$). Thus, participants did become more internally consistent in their categorization (higher β weights) in the Maintenance Phase, but many were becoming consistent in a unidimensional way.

In sum, although our listeners certainly learned to use both dimensions, they did so with considerable difficulty. Also, they tended to use formant frequency more strongly than duration, as indicated by the higher β weights for formant frequency. This is shown in Figure 5 by the strong tendency of the listeners to fall along a line steeper than 45°. Why might listeners rely more on the dimension that is then often abandoned when the Maintenance Phase is uniformly distributed? As described previously, it may be that duration is more salient or easier to encode than formant frequency and that successful learners actively direct their attention to the less salient dimension, overcorrecting for the salience of duration. Recall that a similar pattern was found between participants in the two unidimensional conditions: Participants learned to use formant frequency (when it was relevant) more reliably than duration (when it was relevant), but tended to shift toward using duration in the Maintenance Phase (see Table 4).

Learning a multidimensional category distinction with supervision was difficult but possible, with about half of the participants learning successfully. The analysis of percentage correct and d' data did show a learning effect as did the development in Φ. The consistency measure A did not increase significantly from the first learning phase to the second. The change in both Φ and A in the

[2] Correction for multiple t tests did not substantially alter the results.

[3] The difference between the first learning phase and the Maintenance Phase was marginally significant at $p < .06$.

Table 5
Mean Values and Standard Deviations of the Polar Coordinates φ and A of the β Weights for Duration and Formant Frequency in the Three Phases of Condition 3 and Experiment 1B as Well as the Numbers of Participants Using Only Duration, Only Formant Frequency, or Both

	Condition 3 (Maintenance with equidistant grid)					Experiment 1B (Maintenance with learning stimuli)				
	Φ (σ)	A (σ)	D	Φ	Multi	Φ (σ)	A (σ)	D	Φ	Multi
Learning Phase 1	$n = 6$					$n = 7$				
	0.26 (0.12)	0.21 (0.10)	3	0	3	0.30 (0.09)	0.29 (0.14)	2	4	1
Learning Phase 2	$n = 8$					$n = 8$				
	0.32 (0.18)	0.34 (0.13)	1	1	6	0.37 (0.03)	0.18 (0.21)	0	1	7
Maintenance Phase	$n = 12$					$n = 7$				
	−0.22 (0.31)	0.76 (0.29)	8	0	4	0.24 (0.34)	0.42 (0.18)	0	0	7

Note. Participants using no dimensions significantly are not shown.

Maintenance Phase showed that learning was fragile. Confronted with the equidistantly spaced grid, most listeners opted for a unidimensional solution instead of the multidimensional solution suggested by their prior experience; half of the participants used both dimensions significantly during the last learning phase but only four of them retained this ability in the Maintenance Phase, and the remainder began using duration exclusively.

Experiment 1B addressed two possible explanations for participants' change in categorization strategies when they reached the Maintenance Phase in Condition 3: the absence of feedback in the maintenance phase and the absence of distributional information. If exposure to a uniform distribution of category exemplars as that in Condition 3 is responsible for the altered performance in the Maintenance Phase, performance in this phase should be better when the training distributions are not replaced by the equidistantly spaced grid but instead are maintained. Alternatively, if the absence of trial-by-trial feedback is in itself enough to disturb the previously learned category boundaries, Maintenance Phase performance may be degraded both in Condition 3 and in Experiment 1B.

Experiment 1B: Maintaining a Category Boundary With Distributional Information

Method

Participants. Twelve participants were recruited as in Experiment 1. None had participated in the previous experiment.

Stimuli. As in Condition 3 of Experiment 1 the main axis of variation of the probability density functions was oriented diagonally (see the upper-rightmost panel of Figure 2). Contrary to Experiment 1 however, the maintenance phase stimuli were identical to those in the learning phase (see the lower-rightmost panel of Figure 2). The stimulus characteristics are identical to those for Condition 3 of Experiment 1 (see Table 1).

Procedure. The procedure was identical to that of Condition 3 of Experiment 1. Note that participants did not receive feedback during the maintenance phase. In the maintenance phase, the 112 learning stimuli from each category (224 stimuli) were presented once in random order.

Results and Discussion

Signal detection analysis (percentage correct and d′). There were *t* tests that confirmed that the percentage correct significantly exceeded 50%, and that d' significantly exceeded 0, in both learning phases (all $p < .05$, Bonferroni corrected for multiple comparisons). The bottom row of Table 3 shows these statistics for both conditions.

Although mean performance improved from Learning Phase 1 to Phase 2, as in Condition 3 of Experiment 1, this improvement did not reach significance here, $F(1, 11) = 2.41$, $\eta_p^2 = 0.18$ for percentage correct, $F(1, 11) = 3.24$, $\eta_p^2 = 0.23$ for d'; both *ns*. The magnitude of the improvements in performance was very similar from Condition 3 to the present experiment, as shown in Table 3, and an ANOVA with experiment (Condition 3 and Experiment 1B) and part of the experiment (Learning Phases 1 and 2) yielded no significant main effects or interactions. Thus, differences in participants' performance in the maintenance phase of Experiment 1B and Condition 3 are not attributable to variation in training phase performance.

Logistic regression. As in Experiment 1, the β weights were analyzed using logistic regression. The right side of Figure 5 presents the β weights for duration and formant frequency for each listener in each part of Experiment 1B.

The right-hand column of Figure 5 displays the β weights of each listener in the Formant Frequency × Duration plane for Experiment 1B. The increase in the number of asterisks shows that here, as in Condition 3 of Experiment 1, some listeners learned to use both dimensions in the first Learning Phase, and that performance improved on this measure in the second Learning Phase. In the maintenance phase this learning was maintained among those who had acquired it, contrary to the drop in performance in Condition 3. As a matter of fact, those participants using any dimensions significantly in the Maintenance Phase all used both dimensions.

The right-hand column of Table 5 lists the values of Φ for each phase of the experiment for all listeners who in a given phase used one or more dimensions above chance levels. The mean Φ of the first and second part of the Learning Phase differed significantly from 0, $t_{min} = 8.60$, $p < .05$; and from $(1/2)\pi$, $t_{min} = -0.854$, $p < .05$. Mean Φ values exceeded $(1/4)\pi$ (the value that would reflect an unbiased use of duration and formant frequency), indicating, as in Experiment 1, a stronger reliance on the frequency than on duration.

The analysis of the Maintenance Phase is complicated by an outlier in the lower left quadrant. With the outlier included, Φ was marginally significantly different, $t(7) = 1.98$, $p < .09$

from 0 (duration) and from (1/2)π (formant frequency), $t(7) = -2.19$, $p < .07$. With the outlier collapsed to the upper right quadrant (on the reasonable assumption that the learner retained his or her knowledge of the categories, but inverted the category assignments), mean Φ rose from 0.24 to 0.36, reflecting a preference for formant frequency also observed in the learning phases. In this analysis, mean Φ was significantly different from both 0, $t(7) = 2.37$, $p < .01$ and from (1/2)π, $t(7) = -3.59$, $p < .01$.[4]

The consistency measure *A* showed no significant effect of part of the experiment, $F(2, 10) = 0.82$, *ns*, $\eta_p^2 = 0.14$. Pairwise comparisons showed the difference between the first and second Learning Phases to approach significance ($p < .06$). This did not hold for the differences between each of the Learning Phases and the Maintenance Phase.

These results differ from those of Condition 3 in Experiment 1, in which consistent maintenance of learning was not found, and in which many participants shifted to using duration. This difference between the two Maintenance Phases was significant in an ANOVA with experiment (equidistant grid versus distributional maintenance phase) as a between-subjects factor and *A* as the dependent variable, $F(1, 22) = 18.24$, $p < .05$, $\eta_p^2 = 0.45$; but failed to reach significance when Φ was the dependent variable, $F(1, 18) = 3.03$, $p < .09$, $\eta_p^2 = 0.15$.

In a final analysis, we examined the learning phases of the unidimensional and multidimensional categorization problems, to determine whether the latter categorization was significantly more difficult to learn than the former.[5] An ANOVA with part of the experiment as a within-subjects factor and category structure (unidimensional vs. multidimensional) as between-subjects factor showed that learning a multidimensional distinction was significantly more difficult than learning a unidimensional distinction, both in terms of percentage correct, $F(1, 64) = 144.10$, $p < .05$, $\eta_p^2 = 0.76$; and d', $F(1, 46) = 104.81$, $p < .05$, $\eta_p^2 = 0.69$.

Taken together, the results of Condition 3 of Experiment 1, and Experiment 1B, showed that learning a multidimensional category distinction with supervision was difficult but possible, with about half of the participants learning successfully. Multidimensional learning was fragile and dependent on the continued presence of distributional information. Without it (Condition 3's Maintenance Phase) most listeners opted for a unidimensional solution instead of the multidimensional solution suggested by their prior experience. With distributional information in the Maintenance Phase (Experiment 1B), listeners were able to maintain the use of both dimensions consistently, provided that they had begun to use both in the learning phase.

Experiment 2 investigates unsupervised learning of multidimensionally varying categories. In Conditions 1 and 2 the categorization problems could be solved by using one dimension, whereas the problem presented in Condition 3 required the use of both dimensions. Listeners did not receive any trial-by-trial feedback on their categorization.

Experiment 2: Unsupervised Learning

Method

Participants. Thirty-six (12 in each condition) participants were recruited as in Experiment 1. None had participated in the previous experiments.

Stimuli. The stimuli were identical to those used in Experiment 1: inharmonic sound complexes that varied along the frequency of the spectral peak at which the inharmonic complex was filtered and the duration of the stimulus. See Tables 1 and 2 for detailed stimulus characteristics.

As in Experiment 1, Conditions 1 and 2 differed solely in the relevant dimension of variation, although in Condition 3 both dimensions exhibited relevant variation (see the first three upper panels of Figure 2).

The stimuli in Conditions 1 and 2 manifested relevant variation in one dimension and irrelevant variation in the other so that solving the categorization problem required the use of one dimension only. In Condition 3, the main axis of variation was oriented diagonally as in Condition 3 of Experiment 1 (see the third upper panel of Figure 2).

The maintenance phase of all conditions was identical: Listeners categorized stimuli from an equidistant continuum (see the first three lower panels of Figure 2) as belonging to either group A or B (see also Tables 1 and 2).

Procedure. The procedure was identical to that of Experiment 1, but without trial-by-trial feedback. In the learning phase, listeners heard 448 stimuli (2 categories × 2 repetitions × 112 stimuli per category). Their task was to assign each stimulus to group A or B, using the two-key button box. Once participants had selected a category label on a trial, the monitor would display (the Dutch equivalent of) "next" for 700 ms and the next stimulus was played after a 200-ms blank screen. In the early trials, of course, participants had no basis for choosing one response button rather than the other; over time, as they began to deduce the category structures, their interpretation of the category structures provided them with a way to be consistent in assigning sounds to responses. In the maintenance phase the task was to categorize the sounds from the maintenance continuum. Again, no trial-by-trial feedback was provided.

Results and Discussion

Signal detection analyses. The two upper rows of Table 6 list the percentages correct and d's as well as their standard deviations for the two learning phases of all conditions.

In all conditions, d' exceeded zero for both learning phases, $t_{min} = 2.7$. It should be noted, however, that in Condition 3, in which performance was obviously not as good as in Condition 1 and 2, d' never reached the value traditionally associated with good performance in psychophysical experiments (a d' of 1).

To test whether percentage correct differed from chance, we first calculated the chance level, which is not equal to 50% in an unsupervised learning paradigm. When there is feedback, the mapping of a response to a category can be done a priori (it's possible to be wrong and hence get negative feedback, even on the first response) and the percentage correct can be calculated ac-

[4] Removing the outlier entirely also yielded a significant difference between mean Φ from both 0π, $t(6) = 40.03$, $p < .01$ and (1/2)π, $t(6) = -16.01$, $p < .01$.

[5] This analysis is computed over percentage correct and d', but not the βs (and hence not the Maintenance Phases). Because the multidimensional and undimensional problems yield different optimal numbers of relevant β weights, they are not directly comparable.

Table 6
Signal Detection Results (Mean Percentage Correct and d′) for Experiment 2

	Learning Phase 1		Learning Phase 2	
Variable	% correct (σ)	d' (σ)	% correct (σ)	d' (σ)
Condition 1	0.67 (0.17)	0.78 (0.88)	0.76 (0.20)	1.36 (1.16)
Condition 2	0.62 (0.13)	0.52 (0.65)	0.71 (0.20)	0.99 (1.04)
Condition 3	0.57 (0.05)	0.24 (0.20)	0.59 (0.05)	0.34 (0.19)

cordingly. Without feedback, however, the mapping of the listener has to be inferred based on his or her categorization performance: If he or she tends to use the left button for category A, then that button is assigned to the normative A category and the other button to B. A side effect of this procedure is that listeners always perform at or above the traditional chance level of 50%, even if they are purely guessing, so 50% is not an appropriate chance baseline. Based on the binomial distribution, the expected value of percentage correct for a guessing participant given the current number of trials is 52.66%. Using this value as chance performance, statistical analysis of percentage correct yields results similar to the analyses of d', exceeding chance level in all learning phases, $t_{min} = 2.47$.

To investigate the effect of learning over time, d' and percentage correct were entered into an ANOVA with part of the experiment as a within-subjects variable and condition as a between-subjects variable. For d', there was a significant main effect of part of the experiment, $F(1, 33) = 9.88$, $p < .05$, $\eta_p^2 = 0.23$, indicating a higher d' in the second Learning Phase compared to the d' of the first Learning Phase. Percentage correct showed a similar increase from the first Learning Phase to the second, $F(1, 33) = 8.43$, $p < .05$, $\eta_p^2 = 0.20$. A main effect of condition, $F(2, 33) = 3.47$, $p < .05$, $\eta_p^2 = 0.17$ for percentage correct; and $F(2, 33) = 3.77$, $p < .05$, $\eta_p^2 = 0.19$, for d' indicated the advantage of learning a unidimensional distinction over learning a multidimensional distinction for unsupervised learning. Multiple comparisons confirmed this was due to the difference between unidimensional and multidimensional learning: Condition 1 and 2 did not differ from each other, although both differed significantly from Condition 3.

The above results indicate that even in the absence of trial-by-trial feedback listeners were able to exploit the distributional information available to them.

Logistic regression. As in Experiment 1, we computed a logistic regression analysis with and without the additional interaction term. Of the 108 analyses of Experiment 2 (3 conditions × 3 parts × 12 listeners) only 9 had a significant interaction term. Moreover, the fits of the models with an interaction term were very similar to the fits without an interaction term.

Despite the absence of feedback, most listeners learned to use the relevant dimension. The mean β weight of the relevant dimension was consistently higher than that of the irrelevant dimensions (see Figure 6 and Table 7). The low mean β weights for the irrelevant dimensions indicate that listeners not only learned to use the relevant dimension, but also learned to ignore the irrelevant dimension in the learning phase. The higher β weights of the relevant dimension in Condition 2 (formant frequency relevant) compared to those of Condition 1 (duration relevant) suggest that formant frequency was preferred in learning.

In the Maintenance Phase, listeners had to categorize stimuli from the equidistant grid. Here, the β weight for the relevant dimension dropped in Condition 2 (formant frequency relevant), but not in Condition 1 (duration relevant). Also, in the maintenance phase of Condition 2, listeners started using the irrelevant dimension, duration, much more than in the Maintenance Phase of Condition 1, in which formant frequency was the irrelevant dimension. Similar effects were found in Experiment 1, indicating that the learning advantage for frequency and the bias toward duration in the Maintenance Phase, are not specific to supervised learning.

An ANOVA with dimension (relevant vs. irrelevant) and part of the experiment (Learning Phase 1, Learning Phase 2, Maintenance Phase) as within-subjects factors and Condition (duration relevant

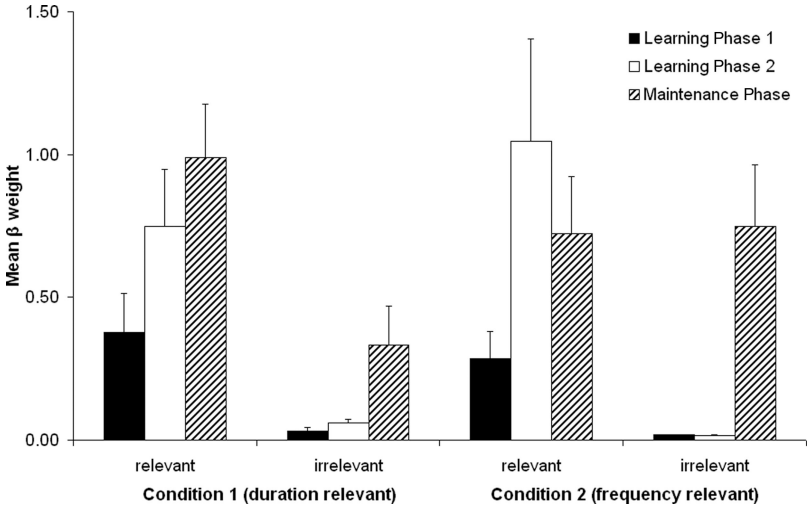

Figure 6. Mean β weights of the relevant and irrelevant dimensions for Condition 1 (duration relevant) and Condition 2 (formant frequency) for each part of Experiment 2.

Table 7
Logistic Regression Results of Condition 1 and 2 of Experiment 2 for Each Condition

	Condition 1 (duration relevant)				Condition 2 (frequency relevant)			
Variable	μ (ß)	σ (ß)	Unidimensional	Multidimensional	μ (ß)	σ (ß)	Unidimensional	Multidimensional
Learning Phase 1								
Relevant	0.38	0.46	5	0	0.47	0.32	6	0
Irrelevant	0.03	0.03	0		0.02	0.01	0	
Learning Phase 2								
Relevant	0.75	0.70	7	0	1.03	1.25	6	0
Irrelevant	0.06	0.05	0		0.02	0.01	0	
Maintenance Phase								
Relevant	0.98	0.65	9	1	0.75	0.69	5	3
Irrelevant	0.33	0.47	2		0.75	0.74	4	

Note. Mean β weights are shown for both dimensions and the number of listeners out of 12 using one (unidimensional) or both (multidimensional) dimensions significantly.

vs. formant frequency relevant) as a between-subjects factor revealed a significant effect of dimension, $F(1, 22) = 25.17$, $p < .05$, $\eta_p^2 = 0.53$, on the β weights, confirming that listeners relied more on the relevant than on the irrelevant dimension. Improvement in categorization over the course of the learning phase was confirmed by a significant main effect of part of the experiment, $F(1, 22) = 18.79$, $p < .05$, $\eta_p^2 = 0.46$). Pairwise comparisons showed each part to differ from each other part ($p < .05$). Condition (duration vs. formant frequency relevant) was not involved in any significant main effects or interactions, although the overall pattern of β weights revealed trends similar to the pattern found in Experiment 1, particularly in the tendency of participants to use duration in the Maintenance Phase even when it was not relevant to the categories learned (see Figure 6).

The results from these two conditions showed that learning of a unidimensional category distinction is possible without the aid of supervision. With duration as the relevant dimension, listeners had no problem categorizing the maintenance stimuli according to the learning distributions. When formant frequency was the relevant dimension, listeners were much more sensitive to the distributional properties of the Maintenance Phase and started using duration more compared to Condition 1. This difference between formant frequency and duration could be due to noisier encoding of formant frequency. Listeners' choice of the relevant dimension based solely on distributional information is an extraordinary feat. This sensitivity to distributional information has rarely been shown in the auditory domain.

Condition 3 investigated learning of a multidimensional category structure with two relevant dimensions of variation. Listeners had to learn a multidimensional distinction: To obtain a high percentage correct, both duration and formant frequency had to be used in the categorization. Figure 7 and Table 8 display the results of Condition 3. Figure 7 plots the β weights of duration and frequency against one another (note that these are the raw β weights, for which the buttons have not been recoded). Participants in the upper-right and lower-left quadrant can be considered to correctly classify the stimuli (assuming their positioning in those quadrants is not due to chance). The only difference between these two quadrants is the reversal in assignment of buttons to categories. In Table 8, the columns on the right-hand side display the number of listeners using a given dimension. The raw data indicate some sensitivity in our listeners to the distributional properties of the stimuli as the data in Table 9 show an increase in the use of both dimensions.

To analyze the multidimensional results presented in Figure 7, we transformed the β weights to polar coordinates as described in Experiment 1. Whether a participant falls into the second quadrant or the third in Figure 7 depended only on which category they

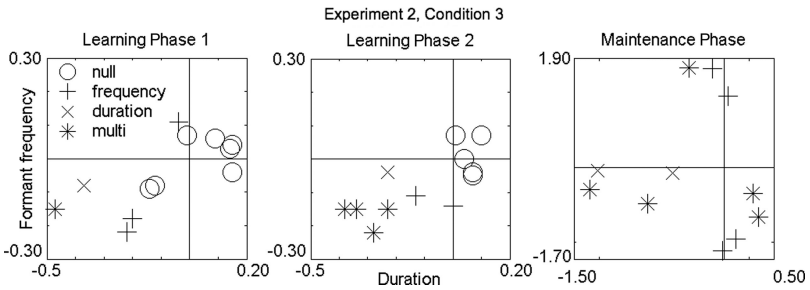

Figure 7. The listeners' β weights associated with the duration and formant-frequency dimensions for the two-dimensional categorization problem of Experiment 2 (Condition 3), plotted in a two-dimensional duration-frequency plane. Asterisks indicate subjects who used both dimensions significantly, Xs indicate listeners who used only duration significantly, plusses indicate listeners who used formant frequency significantly, and listeners marked by circles did not use any dimension significantly.

Table 8
Mean Values and Standard Deviations of the Polar Coordinates φ and A of the β Weights for Duration and Formant Frequency in the Three Phases of Condition 3 of Experiment 2 (Multidimensional Learning) as Well as the Numbers of Participants Using A Only Duration, Only Formant Frequency, or Both

Variable	Φ (σ)	*A* (σ)	Duration	Frequency	Multidimensional
Learning Phase 1 ($n = 6$)	0.25 (0.21)	0.31 (0.14)	2	3	1
Learning Phase 2 ($n = 7$)	0.20 (0.14)	0.28 (0.10)	1	2	4
Maintenance Phase ($n = 12$)	0.20 (0.35)	1.20 (0.42)	5	2	5

Note. Participants using no dimensions significantly are not shown.

assigned to which response button (i.e., two negative βs are the result of mirroring the stimulus assignment as prescribed in Experiment 1) and was therefore arbitrary. For the purpose of analysis, we recoded the Φ values in the lower-left quadrant to Φ values in the upper-right quadrant. The left columns of Table 8 display these recoded mean polar coordinates for each phase of the experiment.

We tested whether the values of Φ differed significantly from the two purely unidimensional solutions (represented by Φs of 0 and (1/2)π). In the first Learning Phase, there was too much variation for mean Φ to significantly differ from either 0, $t(5) = 3.022$., *ns*; or from (1/2)π, $t(5) = –3.27$, *ns*. In the second Learning Phase, however, mean Φ differed significantly from both 0, $t(6) = 3.76$, $p < .051$, and from (1/2)π, $t(6) = –5.64$, $p < .05$. Hence, (some of the) listeners did learn to categorize using both dimensions in the learning phases. In the Maintenance Phase, the mean Φ differed significantly from (1/2)π, $t(11) = –2.95$, $p < .05$; but not from 0, $t(11) = 1.99$, *ns*, reflecting the now familiar preference for duration in the Maintenance Phase.

An ANOVA with *A* as the dependent variable and part of the experiment as a within-subjects variable indicated that listeners as a whole did not get more internally consistent over time as there was no significant effect of part of the experiment, $F(1, 22) = 1{,}68$, *ns,* $\eta_p^2 = 0.02$.

In sum, Experiment 2 showed it to be possible for listeners to benefit from distributional information when learning a multidimensional category distinction. Although not all measures reflected multidimensional learning, listeners were certainly sensitive to the distributional information in the stimuli, both in the signal detection theoretic measures and the mean β weights (as expressed in Φ).

Comparing Supervised and Unsupervised Learning

Unidimensional categorization. Table 9 shows the difference scores of both unidimensional conditions (Experiment 1, Conditions 1 and 2 and Experiment 2, Conditions 1 and 2) in the supervised and unsupervised learning experiments (supervised minus unsupervised for each performance measure). An overall ANOVA with the signal detection measures (percentage correct and *d′*) as dependent variables and learning mode (supervised vs. unsupervised) and condition (duration relevant vs. formant frequency relevant) as independent between-subject measures and part of the experiment as within-subject measure indicated supervised learning to be superior for both percentage correct, $F(1, 44) = 20.14$, $p < .05$, $\eta_p^2 = 0.31$; and *d′* measures, $F(1, 44) = 18.26$, $p < .05$, $\eta_p^2 = 0.29$. Subsequent tests comparing supervised and unsupervised learning in each learning phase separately revealed supervised learning to be superior to unsupervised learning in both learning phases for both conditions for both dependent measures (with Bonferroni correction, the difference in *d′* in the first learning phase was marginally significant [$p < .08$] when

Table 9
Difference Scores of the Unidimensional Supervised (Experiment 1, Conditions 1 and 2) and Unsupervised (Experiment 2, Conditions 1 and 2) Experiment

	Duration relevant			Frequency relevant		
Variable	μ (ß)	% correct	*d′*	μ (ß)	% correct	*d′*
Learning Phase 1						
Relevant	0.28	0.14	0.61	1.08	0.18	0.80
Irrelevant	0.02			0.00		
Learning Phase 2						
Relevant	0.75	0.17	1.23	1.23	0.18	1.08
Irrelevant	0.04			0.02		
Maintenance Phase						
Relevant		0.55			0.51	
Irrelevant		0.23			0.67	

Note. β weights are shown for both dimensions as well as the signal detection analysis measures for the two learning phases. Positive values indicate an advantage for supervised learning; negative values would have indicated an advantage for unsupervised learning (but do not occur due to the superiority of supervised learning).

duration was the relevant dimension, as was the percentage correct [$p < .07$] in the second learning phase when frequency was the relevant dimension).

The effect of supervision on the β weights was investigated in the unidimensional conditions with an ANOVA with part of the experiment (Learning Phase 1, Learning Phase 2, and Maintenance Phase) and dimension (relevant vs. irrelevant) as within-subjects variables and condition (duration relevant vs. formant frequency relevant) and learning mode (supervised vs. unsupervised) as between-subjects factors. This analysis showed a significant advantage for supervised over unsupervised learning, $F(1, 44) = 9.56$, $p < .05$, $\eta_p^2 = 0.18$. Separate analyses per category structure were warranted by the significant three-way interaction between part of the experiment, learning mode and condition. Again, there was an advantage of supervised learning, as evidenced by an effect of learning when duration was the relevant dimension, $F(1, 22) = 5.07$, $p < .05$, $\eta_p^2 = 0.19$, as well as when formant frequency was the relevant dimension, $F(1, 22) = 4.51$, $p < .05$, $\eta_p^2 = 0.17$. The conditions differed solely in the significant interaction between learning mode and part of the experiment that was significant when frequency was the relevant dimension, $F(2, 44) = 17.14$, $p < .05$, $\eta_p^2 = 0.44$, but not when duration was the relevant dimension, $F(2, 44) = 2.17$, *ns,* $\eta_p^2 = 0.09$. Listeners experienced difficulty in the Maintenance Phase in both learning modes when frequency was the relevant dimension. With supervised learning, maintaining formant frequency as the relevant dimension was difficult, whereas with unsupervised learning, it was difficult to suppress the irrelevant dimension duration in the Maintenance Phase.

Separate analyses per learning phase and per condition show a similar picture. In the first learning phase there is a clear effect of supervision when frequency is the relevant dimension, whereas when duration is the relevant dimension, the difference between supervised and unsupervised learning only emerges in the second learning phase. In the Maintenance Phase, when frequency was the relevant dimension, listeners did not use the relevant dimension more than the irrelevant dimension. There was, however, a significant difference between supervised and unsupervised learning. With supervision in the learning phase, listeners experienced difficulty maintaining their use of formant frequency (the β weight for formant frequency decreased), whereas without supervision, they had difficulty suppressing the use of the irrelevant dimensional duration (the β weight for duration increased). In the Maintenance Phase when duration was the relevant dimension, a significant interaction between dimension and learning mode revealed that in supervised learning participants were able to suppress the use of formant frequency in the Maintenance Phase, whereas they could not do so with unsupervised learning. Maintaining the previously learned use of duration proved difficult only in unsupervised learning.

Multidimensional categorization. Multidimensional supervised learning was compared with multidimensional unsupervised learning by comparing the difference scores for percentage correct, d' and the consistency measure A from the logistic regression (see Table 10). With percentage correct as dependent measure, there was a significant advantage for supervised learning in an ANOVA with part of the experiment as within-subjects variable and learning mode (supervised learning vs. unsupervised learning) as between-subject variable, $F(1, 22) = 4.98$, $p < .05$, $\eta_p^2 = 0.19$.

Table 10
Difference Scores of the Multidimensional Supervised (Condition 3 of Experiment 1) and Unsupervised (Experiment 2, Condition 3) Experiment

Variable	% correct	d'	A
Learning Phase 1	0.02	0.09	−0.10
Learning Phase 2	0.02	0.16	0.06
Maintenance phase			−0.44

Note. Signal detection analysis measures are shown for the two learning phases and A is shown for all three phases of the experiment. Positive values indicate an advantage for supervised learning.

This effect was marginal in the analysis of d', $F(1, 22) = 3.55$, $p < .07$, $\eta_p^2 = 0.14$. A similar ANOVA (on only the learning phases) with the consistency measure A as dependent measure also did not reveal a significant difference between supervised and unsupervised multidimensional learning, $F(1, 22) = 0.04$, *ns,* $\eta_p^2 = 0.00$. Post hoc comparisons reveal a significant difference between supervised and unsupervised learning in the Maintenance Phase, with a larger A in the unsupervised condition. This most likely reflects the realization of participants in the supervised condition that a unidimensional solution is incorrect, whereas not all participants in the unsupervised condition may have realized this. This results in a stronger reversal to unidimensional strategies in the Maintenance Phase, which are likely to have a larger consistency measure. The lack of large differences in performance between supervised and unsupervised learning in the learning phases of truly multidimensional categorization problems points to a procedural learning mechanism for multidimensional distinctions—a mechanism that does not depend as much on feedback as verbal or explicit learning mechanisms do.

General Discussion

Listeners provided with trial-by-trial feedback readily learned to differentiate two novel auditory categories that could be distinguished by a single auditory dimension (duration or formant frequency) despite irrelevant variation in the other dimension (Conditions 1 and 2). Learning a truly multidimensional auditory categorization (Condition 3), on the other hand, proved relatively difficult.

Participants' success in generalizing to a Maintenance Phase without supervision depended on whether the relevant dimension was formant frequency or duration, possibly a reflection of processing differences between prothetic or metathetic dimensions (Smits et al., 2006; Stevens & Galanter, 1957) or differences in participants' ability to extract estimates of duration and of formant frequency from the inharmonic complexes used as stimuli. If the categorization problem was truly multidimensional, performance in the Maintenance Phase also depended on whether the stimuli still contained distributional information (Experiment 1B). If the stimuli in the Maintenance Phase lacked distributional information, many participants quickly left their learned multidimensional strategy and reverted to a unidimensional solution, using the dimension of duration.

The results of Experiment 2 make it clear that unsupervised learning of multidimensional auditory categories is feasible. Lis-

teners are thus remarkably sensitive to distributional information. Given two equally salient dimensions with an equal amount of variation, listeners will still use the dimension with relevant variation in their categorizations. There were important differences between the learning of unidimensional category distinctions and multidimensional category distinctions as well as between supervised and unsupervised learning.

With only one relevant dimension of variation (Condition 1 and 2), unsupervised learning was surprisingly good in the learning phase, despite the absence of trial-by-trial feedback. The robustness of this learning depended largely on which dimension was the relevant one. When duration was the relevant dimension, most listeners were able to generalize their successful categorization strategy to the Maintenance Phase, in which distributional cues were no longer present. When formant frequency was the relevant dimension, listeners found it much more difficult to suppress the use of the irrelevant dimension duration in the Maintenance Phase. The emerging use of the irrelevant dimension in the maintenance phase in both conditions of unidimensional learning can be interpreted as a loss of previously learned category distinctions, but also can be considered as evidence of the sensitivity of listeners to the absence of the distributional cues that had been present in the Learning Phase.

When there were two relevant dimensions of variation (Condition 3), learning to use both dimensions to correctly categorize the stimuli was much more difficult, but there was not as much difference between supervised and unsupervised learning as was found for unidimensional category learning problems. Listeners were clearly sensitive to the distributional information present in the stimuli, but not all reached a suitable categorization strategy during the 440 learning stimuli. It might be that there were not enough trials to show a larger learning effect (and the marginally significant results do suggest so), but the absence of a difference in the learning phases suggests that such learning is slow, at best.

Performance tended to decline in the maintenance phase in which stimuli were drawn from a uniform distribution and presented without feedback. There are several possible explanations for this.

First, the testing of new tokens per se, and not the distributional characteristics of those new tokens, may have led to changed performance on the uniform grid. We consider this unlikely because of the large number of category exemplars (224) that were each presented only twice during training. Participants probably did not learn to respond to only the set of trained exemplars themselves; rather, they learned to respond to the categories, with a response strategy presumably generalizable over similar novel exemplars (see also Greenspan et al., 1988, for the importance of sufficient variation in the stimuli). Second, in the absence of trial-by-trial feedback participants may have simply "started over," noting the change in the procedure and putting less weight on their previous learning, while at the same time becoming attuned to the novel distribution of stimuli. Many of the test stimuli fell in the region between the trained categories. Such exposure in sufficient quantity should count as evidence to the learner that in fact the two categories are one and the same, for precisely the same reason that distributional learning of categories was possible in the first place. What counts as a "sufficient quantity" should depend on how readily the learner allows new evidence to override earlier, well-supported assumptions. Third, the relatively restricted range of stimulus values in the Maintenance Phase may have contributed to the disappearance of multidimensional categorization in that phase. It is conceivable that the more extreme stimuli of the learning phase "anchored" participants' memory representations of the dimensions of variation, particularly for formant frequency, and once this variation was reduced, they had more difficulty recovering frequency information from the maintenance stimuli. This hypothesis could be examined by testing performance on a broader grid of maintenance stimuli.

The use of a grid with equidistantly spaced stimuli to assess the psychophysical space of a listener is a standard technique. The lack of information in the distribution of the stimuli is intended to neutrally probe the participants' psychophysical space and prevent participants from changing their categorization tendencies. However, this is not what happened in our experiments; our listeners picked up on the fact that in the Maintenance Phase the category structure was no longer present, and altered their categorizations. When continuously confronted with stimuli that contained distributional information, their performance level hardly dropped when feedback was discontinued. These discrepancies warrant further research into the robustness of auditory and visual category learning. This result has implications for speech research that uses similar equidistant continua to investigate newly established speech contrasts (Repp & Liberman, 1987), which might be susceptible to rapid degradation resulting from the lack of distributional information at test.

Both Experiments 1 and 2 evidenced a differential effect of dimension, particularly in the maintenance phase in which duration seemed to be the dimension of choice. In the absence of distributional cues, participants were more prone to use duration than formant frequency in their categorization. We have raised the possibility that this preference may be explained in terms of Stevens and Galanter's (1957) distinction between prothetic and metathetic dimensions. Another, more likely, explanation is to extend Ashby et al.'s (1999) distinction between rules that are easy to verbalize and rules that are hard to verbalize. Especially in Condition 3 of Experiment 1, several participants reported being at a loss in the maintenance phase and opting for the duration distinction because it was easier to distinguish the sounds based on duration. Further, when asked about the two dimensions of variation, most participants found it harder to describe the formant frequency dimension than the durational dimension. Formulating a verbal rule in the maintenance phase might be easier with duration as the relevant dimension.

Learning to categorize auditory stimuli with more than one relevant dimension of variation is a task faced by infants and learners of a second language. In the case of infants, and perhaps in many instances of second language learning, this is an unsupervised learning process. Recent studies have suggested that under some circumstances infants can learn unidimensional speech categories without feedback (Maye et al., 2002), even when given only 96 stimulus exposures. It should be noted that the stimuli in this study only contained one relevant dimension of variation, and did not display irrelevant variation in another dimension. All current theories of infant phonetic category learning assume that infants can compute categories from phonetic distributions; the Maye et al. (2002) result suggested that this learning might in fact be extremely rapid, helping to account for infants' precocious

acquisition of native phonetic categories (e.g., Polka & Werker, 1994).

Although there are obviously a number of important differences between the present study and the infant experiments, we believe it is worth considering the possibility that the category learning mechanisms probed in the present studies are similar to those used by infants to induce the categories of their language. If so, it seems reasonable to suppose that infants, like some of the adults in the present studies, make initial assumptions about phonetic categories that are not accurate, for example by favoring unidimensional solutions to multidimensional phonetic problems, or by showing delayed category learning when the distributional evidence contains tradeoffs among distinct dimensions. Although phonetic cue-trading experiments with infants now have a long history (e.g., Eimas & Miller, 1980), relatively little developmental work has attempted to discover how infants' learning of native-language speech categories is affected by dimensional structure.

This discrepancy between our findings and infants' apparently near-universal success in multidimensional (phonetic) category learning can be addressed in several ways. First, both Ashby et al. (1999) and Love (2002) argued that participants initially opt for a unidimensional solution when they are faced with a new categorization problem. Only when there is sufficient negative feedback will they switch to a multidimensional strategy. Most studies construe this negative feedback as trial-by-trial feedback (Ashby et. al, 1998; Maddox, Ashby & Waldron, 2002; Maddox, Bohil, & Ing, 2004). However, our experiments showed approximately equally poor supervised and unsupervised learning of multidimensional categories. Perhaps learning multidimensional auditory categories is not influenced as much by supervision as learning multidimensional visual categories is. In the approach of Gureckis & Love (2003) trial-by-trial feedback is not necessary, and a "surprising" event can change the categorization behavior of the model. Although our listeners clearly were sensitive to the distributional information in the stimuli, the discrepancy between their categorizations and the probability density functions may not have been surprising or salient enough to prompt a switch to a multidimensional rule.

A second explanation is that infants receive much more exposure than adults did in our experiments. Though Maye et al. (2002) showed that short-term modification of infants' speech categories is possible with very little training, it remains the case that infants' day-to-day exposure to speech dwarfs the 440 stimuli our participants listened to. Indeed, American infants hear between 500 and 1,500 words spoken to them by their parents each hour they interact (Hart & Risley, 1995, p. 239; see also Swingley, 2007). The everyday speech input infants receive, on the other hand, is much more complex in terms of contextual variability and talker characteristics than our stimuli. Hence, it is difficult to compare the relative difficulties of the learning task faced by infants and the one faced by our listeners. Other possibilities are open; for example, perhaps infants are simply better learners than adults, or perhaps the categories we tested here had auditory properties that made them harder to learn than speech categories. Note, though, that whereas the artificial categories we tested can be separated by simple linear boundaries drawn in two-dimensional space, no computational model of any kind has ever successfully induced the phonetic categories of any language from ordinary conversational infant-directed speech (e.g., Lin, 2005; though see Vallabha, McClelland, Pons, Werker, & Amano, 2007).

The comparison of the supervised and unsupervised learning experiments showed an overall advantage for supervised learning. This was especially clear in the unidimensional learning experiments. There, supervision helped suppress the tendency to use the irrelevant dimension in the test phase. Performance in unsupervised learning of a unidimensional category structure was still surprisingly good, considering that listeners' only source of information was the distribution of the stimuli in perceptual space. The large advantage for supervised learning that was found for unidimensional learning was not present for multidimensional learning. There actually was a small advantage for unsupervised learning in the Maintenance Phase, which might have been due to the similar procedure for the training and the maintenance phase in the case of unsupervised learning. With supervised learning, participants were faced with the sudden withdrawal of trial-by-trial feedback in the test phase, possibly causing some confusion, whereas this was not the case in the unsupervised learning experiments.

These experiments show that listeners perform well with categories with only one relevant dimension of variation despite the presence of substantial irrelevant variation. This learning is fragile, judging by the change in categorization behavior of listeners when confronted with stimuli without distributional information. The categories that were most similar to real speech, in that they were defined by truly multidimensional variation, were the hardest for these adult listeners to acquire.

References

Abel, S. M. (1972). Duration discrimination of noise and tone bursts. *Journal of the Acoustical Society of America, 51,* 1219–1223.

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Ainsworth, W. A. (1972). Duration as a cue in the recognition of synthetic vowels. *Journal of the Acoustical Society of America, 51,* 648–651.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology, 39,* 216–233.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105,* 442–481.

Ashby, F. G., & Maddox, W. T. (1990). Integrating information from separable psychological dimensions. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 598–612.

Ashby, F. G., & Maddox, W. T. (1993). Relationships between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology, 37,* 372–400.

Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition, 30,* 666–677.

Ashby, F. G., Queller, S., & Berretty, P. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception and Psychophysics, 61,* 1178–1199.

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin and Review, 6,* 363–378.

Aslin, R. N., Jusczyk, P. W., & Pisoni, D. B. (1998). Speech and auditory processing during infancy: Constraints on and precursors to language. In W. Damon (Series Ed.) & D. Kuhn & R. Siegler (Vol. Eds.), *Handbook of child psychology: Vol. 2: Cognition, perception and language* (5th ed., pp. 147–198). New York: Wiley.

Aslin, R. N., Pisoni, D. B., & Jusczyk, P. W. (1983). Auditory development and speech perception in early infancy. In M. Haith & J. Campos (Eds.),

Handbook of child psychology, infancy and developmental psychobiology (Vol. 2, pp. 573–687). New York: Wiley.

Attneave, F. (1957). Transfer of experience with a class-schema to identification-learning of patterns and shapes. *Journal of Experimental Psychology, 54,* 81–88.

Best, C. T. (1995). A direct realist view of cross language β speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–206). Baltimore: New York Press.

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

Best, C. T., & Strange, W. (1992). Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of Phonetics, 20,* 305–330.

Booij, G. (1995). *The phonology of Dutch.* Oxford, UK: Clarendon.

Bradlow, A. R., Pisoni, D. B., Yamada, R. A., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America, 101,* 2299–2310.

Cameron Marean, G., Werner, L. A., & Kuhl, P. (1992). Vowel categorization by very young infants. *Developmental Psychology, 28,* 163–405.

Egeth, H. E., & Mordkoff, J. T. (1991). Redundancy gain revisited: Evidence for parallel processing of separable dimensions. In J. Pomerantz & G. Lockhead (Eds.), *The perception of structure* (pp. 131–143). Washington, DC: American Psychological Association.

Eimas, P. D., & Miller, J. L. (1980). Contextual effects in infant speech perception. *Science, 209,* 1140–1141.

Feldman, J. (2000). Minimization of Boolean complexity in human category learning. *Nature, 407,* 630–633.

Fiser, J., & Aslin, R. N. (2001). Statistical learning of higher-order temporal structure from visual shape-sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 458–467.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp 233–277). Baltimore: York.

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

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

Francis, A. L., Nusbaum, H. C., & Fenn, K. (2007). Effect of training on the acoustic-phonetic representation of synthetic speech. *Journal of Speech, Language, and Hearing Research, 50,* 1445–1465.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 234–257.

Garner, W. R. (1974). *The processing of information and structure.* Potomac, MD: Erlbaum.

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research, 47,* 103–138.

Goldstone, R. L., & Kersten, A. (2003). Concepts and categorization. In A. F. Healy & R. W. Proctor (Eds.), *Comprehensive handbook of psychology: Vol. 4: Experimental psychology.* (pp. 599–621). Hoboken, NJ: Wiley.

Gottwald, R. L., & Garner, W. R. (1972). Effects of focusing strategy on speeded classification with grouping, filtering, and condensation tasks. *Perception and Psychophysics, 11,* 179–182.

Goudbeek, M., Swingley, D., & Kluender, K. R. (2007, August). *The limits of multidimensional category learning.* Proceedings of Interspeech 2007, Antwerp, Belgium.

Grau, J. W., & Kemler Nelson, D. G. (1988). The distinction between integral and separable dimensions: Evidence for the integrality of pitch and loudness. *Journal of Experimental Psychology: General, 117,* 347–370.

Greenspan, S. L., Nusbaum, H. C., & Pisoni, D. B. (1988). Perceptual learning of synthetic speech by rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 3,* 421–433.

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America, 100,* 1111–1121.

Gureckis, T. M., & Love, B. C. (2003). Human unsupervised and supervised learning as a quantitative distinction. *International Journal of Pattern Recognition and Artificial Intelligence, 17,* 885–901.

Hart, B., & Risley, T. R. (1995). Meaningful differences in the everyday experience of young American children. Baltimore: Brookes.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America, 97,* 3099–3111.

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *Journal of the Acoustical Society of America, 119,* 3059–3071.

Homa, D., & Cultice, J. C. (1984). Role of feedback, category size, and stimulus distortion on the acquisition and utilization of ill-defined categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 83–94.

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

Kemler Nelson, D. G. (1993). Processing integral dimensions: The whole view. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 1105–1113.

Kewley-Port, D., & Watson, C. S. (1994). Formant-frequency discrimination for isolated English vowels. *Journal of the Acoustical Society of America, 95,* 485–496.

Kuhl, P. K. (1985). Categorization of speech by infants. In J. Mehler & R. Fox (Eds.), *Neonate cognition: Beyond the blooming buzzing confusion* (pp. 231–262). Hillsdale, NJ: Erlbaum.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science, 9,* F13–F21.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255,* 606–608.

Lin, Y. (2005). *Learning features and segments from waveforms: A statistical model of early phonological acquisition.* Unpublished doctoral dissertation, University of California at Los Angeles.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/: II. The role of phonetic environment and talker variability in new perceptual categories. *Journal of the Acoustical Society of America, 94,* 1242–1255.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin and Review, 9,* 829–835.

Maddox, W. T., Ashby, F. G., & Waldron, E. T. (2002). Multiple attention systems in perceptual categorization. *Journal of Memory and Language, 30,* 325–339.

Maddox, W. T., Bohil, C. J., & Ing, A. D. (2004). Evidence for a procedural learning-based system in category learning. *Psychonomic Bulletin and Review, 11,* 945–952.

Maddox, W. T., Ing, A. D., & Lauritzen, J. S. (2006). Stimulus modality interacts with category structure in perceptual category learning. *Perception & Psychophysics, 68,* 1176–1190.

Maye, J., & Gerken, L. (2000). Learning phoneme categories without

minimal pairs. *Proceedings of the 24th Annual Boston University Conference on Language Development, 2,* 522–533.

Maye, J., & Gerken, L. (2001). Learning phonemes: How far can the input take us? In A. H.-J. Do, L. Domínguez, & A. Johansen (Eds.*), Proceedings of the 25th annual Boston University conference on language development* (pp 480–490). Somerville, MA: Cascadilla Press.

Maye, J., Werker, J., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82,* B101–B111.

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

McClelland, J. L., Fiez, J. A., & McCandliss, B. D. (2002). Teaching the /r/-/l/ distinction to Japanese adults: Behavioral and neural aspects. *Physiology and Behavior, 77,* 657–662.

Melara, R. D., & Marks, L. E. (1990). Perceptual primacy of dimensions: Support for a model of dimensional interaction. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 398–414.

Nooteboom, S. G., & Doodeman, G. J. N. (1980). Production and perception of vowel length in spoken sentences. *Journal of the Acoustical Society of America, 67,* 276–287.

Nosofsky, R. M. (1990). Exemplar-based approach to categorization, identification and recognition. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 363–393). Hillsdale, NJ: Erlbaum.

Peperkamp, S., Pettinato, M., & Dupoux, E. (2003). Allophonic variation and the acquisition of phoneme categories. In A. Brown & F. Conlin (Eds.), *Proceedings of the 27th annual Boston University conference on language development* (pp. 650–661) Somerville, MA: Cascadilla Press.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of vowels. *Journal of the Acoustical Society of America, 24,* 175–184.

Pierrehumbert, J. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech, 46,* 115–154.

Polivanov, E. (1931). La perception des sons d'une langue étrangère. *Travaux du Cercle Linguistique de Prague, 4,* 79–96.

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of non-native vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 421–435.

Pomerantz, J. R., & Lockhead, G. R. (1991). Perception of structure: An overview. In G. R. Lockhead & J. R. Pomerantz (Eds.), *The perception of structure* (pp. 1–20). Washington, DC: American Psychological Association.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77,* 353–363.

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology, 83,* 304–308.

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin, 82,* 81–110.

Repp, B. H., & Liberman, A. M. (1987). Phonetic category boundaries are flexible. In S. R. Harnad (Ed.), *Categorical perception. The groundwork of cognition* (pp. 89–112). Cambridge, UK: Cambridge University Press.

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

Shepard, R. N. (1991). Integrality versus separability of stimulus dimensions. In G. R. Lockhead & J. R. Pomerantz (Eds.), *The perception of structure* (pp. 53–72). Washington, DC: American Psychological Association.

Smits, R., Sereno, J., & Jongman, A. (2006). Categorization of sounds. *Journal of Experimental Psychology: Human Perception and Performance, 32,* 733–754.

Stevens, S. S., & Galanter, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology, 54,* 377–411.

Swingley, D. (2003). Phonetic detail in the developing lexicon. *Language and Speech, 46,* 265–294.

Swingley, D. (2007). Lexical exposure and word-form encoding in 1.5-year-olds. *Developmental Psychology, 43,* 454–464.

Tyler, M. D., & Johnson, E. K. (2006, June). *Testing the limits of artificial language learning*. Poster presented at the 15th Biennial International Conference on Infant Studies, Kyoto, Japan.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences of the United States of America, 104,* 13273–13278.

Wade, T., & Holt, L. L. (2005). Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *Journal of the Acoustical Society of America, 118,* 2618–2633.

Werker, J. F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., & Amano, S. (2007). Infant-directed speech supports phonetic category learning in English and Japanese. *Cognition, 103,* 147–162.

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.

Younger, B. A. (1985). The segregation of items into categories by ten-month-old infants. *Child Development, 56,* 1574–1583.

Zwicker, E., & Fastl, F. (1990). *Psychoacoustics: Facts and models.* Berlin, Germany: Springer Verlag.

Received October 8, 2007
Revision received August 7, 2008
Accepted October 28, 2008 ■