Chapter 22

ACQUIRING AUDITORY AND PHONETIC CATEGORIES*

MARTIJN GOUDBEEK, ROEL SMITS, AND ANNE CUTLER

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

DANIEL SWINGLEY

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, University of Pennsylvania, PA, USA

Contents

Abstract

Infants' first steps in language acquisition involve learning the relevant contrasts of the language-specific phonemic repertoire. This learning is viewed as the formation of categories in a multidimensional psychophysical space. Categorization research in the visual modality has shown that adults are unable to learn multidimensional categorization problems without supervision. The success of infants in acquiring a phonetic system suggests that formation of multidimensional categories should be more tractable in the auditory modality. We describe experiments investigating adult learning of multidimensional speech and nonspeech categories. These experiments revealed that the degree of difficulty is actually significantly greater than that observed in the visual modality. Despite comparable methods, our results differ from those in visual category learning: feedback that is effective for visual categories is not effective in the auditory modality. Attending to more than one dimension in auditory category formation is possible for adult listeners, but very hard.

* Part of this research was carried out with financial support from the Dutch Scientific Research Council. We further thank Keith Kluender, University of Wisconsin, Madison, for financial and other assistance with the speech experiments.

Handbook of Categorization in Cognitive Science, Edited by Henri Cohen and Claire Lefebvre

1. Introduction

Both infants and learners of a second language are faced with the problem of correctly categorizing the sounds of a new language. This task breaks down into two problems for the listener. The first is how to acquire the categories in the first place. Starting without any knowledge of the sounds of a language, how do listeners extract the categories from the language around them? The second problem is the mapping of incoming acoustic material to the categories, once they have been acquired. In this chapter, we are primarily concerned with the first question. The online mapping and representation of auditory categories is investigated by, for example, Smits, Sereno, and Jongman (in press). Learning the sounds of a second language presents the listener with the additional problem of a psychophysical space that is already structured by the first language [Best, Roberts, and Sithole (1988)]. The basic process of acquiring the categories of a second language is the same as with the first, although the first language will always interfere with listening to a second language [Cutler and Broersma (2005)].

Our approach to the issue of the acquisition of an auditory category is similar to that applied in visual category learning studies [Nosofsky (1992), Ashby and Maddox (1993)]. Analogous to the studies investigating visual category learning, our work concerns perceptual categories, i.e., categorization in a psychophysical space with continuous dimensions. This may be different from categorization learning for "higher-order" categories involving dichotomous features such as "shape of the head" (round versus oval) or "long versus short legs" [Minda and Smith (2001, 2002)]. We assume that when listeners hear a sound, this sound is evaluated in a number of dimensions and mapped onto a point in a multidimensional psychophysical space. Repeated exposure to sounds originating from distinct categories will thus lead to "clouds" of points in the space. If, after a period of exposure, several more or less distinct clouds emerge, the subjects may start to associate the clouds with different categories. Such a conceptualization was implemented by Behnke (1998) in a neural network that recognizes patterns in a phonetic map; Kornai (1998) likewise constructed a neural network modeling the data of Peterson and Barney (1952).

The capacity of newborns to discriminate speech sounds is remarkable. In the first few months of life, they discriminate a wide range of speech-sound contrasts, but over the course of the first year they start to conflate similar sounds if those sounds do not make a phonological contrast in their native language [see, e.g., Aslin, Jusczyk, and Pisoni (1998) or Jusczyk (1997) for reviews]. Several studies have found decrements in non-native consonant discrimination by the age of 12 months [e.g., Werker and Tees (1984)] and analogous decrements in non-native vowel perception even earlier [Kuhl et al. (1992), Polka and Werker (1994)].

The above description holds for the discrimination of speech sounds, but not necessarily for their categorization. Discrimination involves the comparison of at least two recently heard stimuli, but categorization involves the comparison of a stimulus to a stored category representation. Category learning entails that certain stimuli that may be perceptually distinguishable are associated with a single category, and thus become, in some sense, identical. Eventually, such category learning may cause the once easily discriminable stimuli to become less discriminable or even indiscriminable [Goldstone (1994)]. Thus, the learning of categories may at least partly explain the loss of infants' abilities to discriminate between phonemes that are not contrastive in their native language. Infants' phonetic categorization abilities are not simply equivalent to their discriminatory abilities, because true acquisition of a phonetic category requires not simply the failure to discriminate between members of that category, but also the ability to assign the appropriate category label to these members. Perception of similarity, as evidenced, for example, by the studies of Kuhl [Hillenbrand (1983), Kuhl (1985), Cameron Marean, Werner, and Kuhl (1992)], is undoubtedly an important aspect of categorization. However, the literature does not yet, in our opinion, motivate a complete account of the relationship between stimulus discrimination and stimulus categorization by infant listeners.

The acquisition of speech categories by infants happens in a wholly unsupervised way. Simply by being exposed to speech, and without formal instruction, infants acquire their native phonetic categories and lose the ability to discriminate non-native contrasts. This process occurs too early in development to be driven by semantic contrasts in phonetically similar words and is therefore generally considered to result from infants' distributional analysis of speech. Supporting this notion, experimental studies have shown that infants are extremely sensitive to the statistical properties of incoming speech signals [Saffran, Aslin, and Newport (1996), Kuhl (2000), Saffran (2001)]. Of particular relevance to phonetic category learning is a study by Maye, Werker, and Gerken (2002), in which two groups of infants were exposed to /da/–/ta/ stimuli in a preferential looking procedure. One group listened to a unimodal distribution, the other to a bimodal distribution of stimuli, encouraging them to group the sounds into two categories. Infants who listened to a bimodal distribution in the training phases listened longer to Alternating (two different stimuli) trials compared to Non-Alternating (the same stimulus repeated) trials in the test phase. Infants exposed to a unimodal distribution did not show this differential looking [Maye et al. (2002)]. Similar sensitivity to distributional properties has been found for adults [Maye and Gerken (2000, 2001)]. Thus, Maye and colleagues have shown that both infants and adults are capable of learning unidimensional categories without supervision. As Pierrehumbert (2003) points out, however, it is not clear whether this result would generalize to less salient phonetic dimensions or to noisier training conditions.

We assume that statistical learning lies at the heart of phonetic category acquisition; acquiring these categories is simply equivalent to recognizing the statistical patterns present in speech. Statistically based category learning has previously been most intensively studied with respect to visually presented stimuli, and our experimental approach uses methods and insights from such visual categorization studies.

The general methodology is described in Section 2. In Section 3, we describe a study investigating the learning of two nonspeech categorization problems, one with a unidimensional category distinction and one with a multidimensional category distinction. In the third section, the learning of speech categories is investigated, again with a unidimensional and a multidimensional category distinction. The last section summarizes and discusses the presented experiments.

2. Testing category learning

In a two-dimensional psychophysical space we defined two categories as two-dimensional probability density functions, each represented by a "cloud" of points. We varied the relevance of each of the dimensions by manipulating the orientation of the clouds. For example, exposure to the structure in the top left cell in Figure 1 should encourage subjects to categorize using only the vertical dimension (dimension 1), whereas exposure to the structure in the bottom left figure should encourage subjects to use only dimension 2. Exposure to the structures in the right-hand column should increase the tendency to use both dimensions in categorizing, because the use of only one dimension would lead to many incorrect categorizations.

The experiments described in this chapter used a single basic procedure, in which a training phase was followed by a test phase. In the training phase, subjects heard

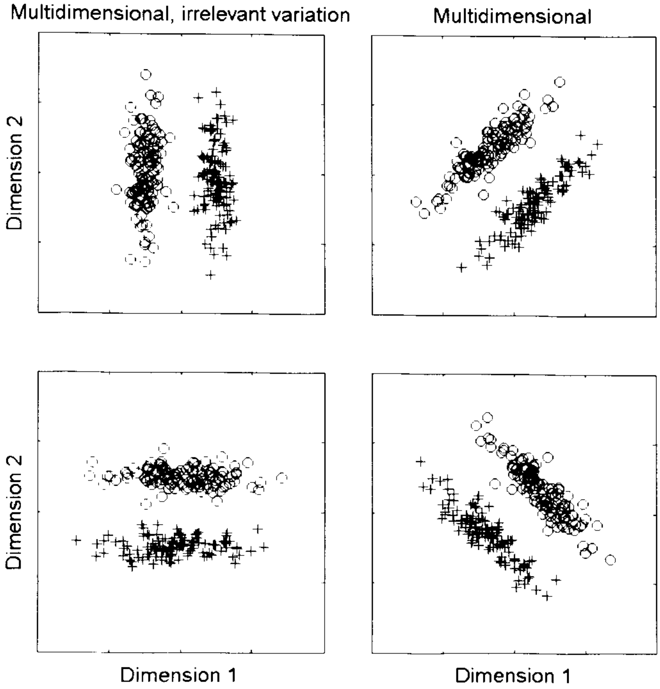

Fig. 1. Four category structures in a two-dimensional space.

stimuli drawn from two probability density functions. Solving the categorization problem required the use either of one dimension or of both dimensions simultaneously, depending on the experiment. If listeners chose a unidimensional criterion, they would assign all stimuli below a criterial value on that dimension to one category and all stimuli above that value to the other category. To categorize via more complex rules, such as the diagonal line that separates the categories in the right-hand column of Figure 1, they would have to use more dimensions. After the training phase, the listeners entered a test phase, which was intended to assess the subjects' categorization behavior across the psychophysical space. Subjects heard and categorized stimuli that were drawn from an uniformly spaced grid, in which distributional information was no longer present (see the right panel of Figure 2).

The experiments used supervision, i.e., feedback, in the training phase. This contrasts with the infant situation where supervision is absent, but allows for a direct comparison with the visual perception studies and models of second language learning in adults; supervised learning of unidimensional and multidimensional auditory categories has seldom been studied and may thus inform models of the learning of second language categories, a process which, at least in part, is often supervised.

One hazard in applying the above methodology to speech perception experiments is that subjects may use the representation of the sounds of their native language in solving the categorization problem [Best and Strange (1992)]. We therefore first presented listeners with nonspeech stimuli that have been shown to be dissimilar to any known sound categories [Smits et al. (in press)]. Combining experiments with nonspeech stimuli as an analog for speech with experiments with speech stimuli also addresses the nature of the exact relationship between speech and nonspeech stimuli. Speech and nonspeech could be analyzed by partly different perceptual systems, but they could also to a large extent be perceived by the same system. In the set of experiments with speech stimuli, we minimized the influence of existing native categories by using vowel stimuli that would map to relatively "empty" areas of the subjects' native vowel space.

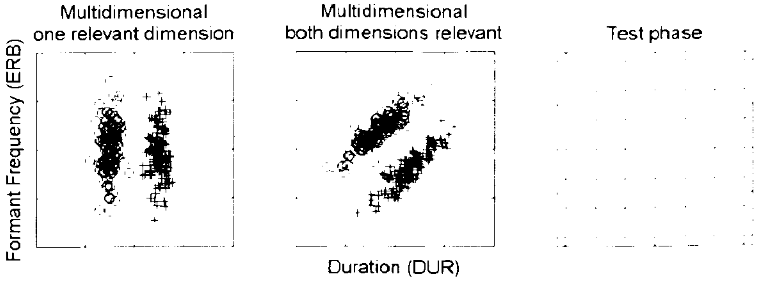

Fig. 2. Category structures of conditions 1 and 2 (left and middle panels) and the test phase (right panel) for both the speech and nonspeech experiment.

3. Learning of nonspeech categories

In our first study, subjects had to learn either a unidimensional (condition 1) or a multidimensional category structure (condition 2). In condition 1, only duration was important for creating the category distinction (see the left-hand panel of Figure 2), and subjects had to learn to ignore the other dimension, whereas in condition 2, both dimensions were relevant (see the middle panel of Figure 2). The category learning process was supervised: subjects received trial-by-trial feedback. In the test phase, listeners categorized stimuli from the equidistantly spaced grid (see the right-hand panel of Figure 2) without receiving feedback.

The stimuli were inharmonic tone complexes filtered by a single resonance, 112 in each category. We used inharmonic sounds because harmonic complexes tend to be easily associated with speech. The two dimensions of variation in the experiment were the frequency of the spectral peak at which the sound complex was filtered ("formant frequency") and the duration of the stimulus ("duration"). Many previous investigations have shown these dimensions to be very important in the perception of vowel sounds [e.g., Ainsworth (1972), Peterson and Barney (1952)]. Detailed descriptions of the stimulus construction can be found in Smits et al. (in press).

Table 1 lists the properties of the stimuli. The distributions were defined in a "perceptual scale" spanned by psychological scales. The psychophysical space commonly accepted for the perception of frequency is the equivalent rectangular bandwidth (ERB) scale [Glasberg and Moore (1990)]. The ERB is derived from frequency according to the following transformation, where f refers to frequency in Hertz:

$$\mathrm{ERB} = 21.4 \times 10 \log(0.00437 * f + 1)$$

Table 1
Distributional characteristics of the stimuli for both training conditions (relevant variation in one dimension and relevant variation in two dimensions) and for the test phase of the nonspeech experiment

	Training stimuli					
	Category A			Category B		
	Means	σ	ρ	Means	σ	ρ
Condition 1	47.7 DUR 18.8 ERB	0.7 DUR 1.88 ERB	−0.05	52.5 DUR 18.9 ERB	0.7 DUR 1.9 ERB	−0.1
Condition 2	48.4 DUR 17.8 ERB	2.8 DUR 1.3 ERB	−0.9	51.7 DUR 19.7 ERB	2.8 DUR 1.3 ERB	−0.9

Test stimuli

Dimension	Mean	Min	Max	Step size
Duration	50.1 DUR	47.6 DUR	52.6 DUR	0.5 DUR/step
Frequency	18.8 ERB	17.6 ERB	20.0 ERB	0.2 ERB/step

Psychological duration D, the psychophysical counterpart of duration, was calculated from stimulus duration t by Smits et al. based on data provided by Abel (1972) as follows:

$$D = 10 \log t$$

To ensure that both dimensions would be equally salient and discriminable, they were normalized using their respective just noticeable differences (Weber fractions). For formant frequency, this is equivalent to 0.12 ERB in the relevant frequency region [Kewley-Port and Watson (1994)], while for duration pilot experiments of Smits et al. (in press) and our subsequent piloting with stimuli varying in duration and frequency indicated that 0.25 D resulted in a discriminability comparable to that of 0.12 ERB. These values were used to establish the range of variation of the stimuli, so that the difference between the lowest and the highest value on both the frequency and the duration dimension equaled 20 just noticeable differences. Twelve subjects, students from the University of Nijmegen without a history of hearing problems, participated in return for a small payment. They were seated in a soundproof booth in front of a computer screen and a response button box. In the training phase, they listened to 448 stimuli, two times 112 from each of the two categories they were attempting to learn. In condition 1, the stimuli varied irrelevantly in one dimension, duration, and relevantly in the other dimension (frequency of the spectral peak). In condition 2, the stimuli manifested relevant variation in both dimensions (see the left and middle panels of Figure 2). The subjects' task was to assign each stimulus to group A or B, using the two-key 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. After the visual feedback disappeared, the next stimulus was presented. Subjects were offered a brief pause halfway through the training phase, and again between the second part of the training and the test phase. In the test phase, subjects categorized sounds from the test continuum (right panel of Figure 2) as belonging to group A or B, this time without receiving feedback regarding their response.

Subjects' responses were analyzed using a logistic regression technique. Figure 3 shows the mean β-weights from the logistic regression separately for the first half of the training ("Training Part 1"), the second half of the training, ("Training Part 2") and the test stimuli ("Test"). The β-weight indicates the extent to which a factor (dimension) can be used to fit the data with a logistic function. We divided the results for the training phase into two parts to probe for learning during the training phase. Table 2 displays the full results of the logistic regression analysis. Subjects clearly learned to use the relevant dimension in the training phase, and they retained this ability in the test phase.

In the individual analysis of each subject, the size of the β-weight may or may not be significantly different from zero. When it is not significantly different from zero, subjects did not make any significant use of the dimension in question. The columns labeled "Uni" and "Multi" show this information. A subject who uses both dimensions is counted under "Multi," while a subject who uses only one dimension to a significant extent is counted under "Uni." Subjects who use no dimension are not counted.

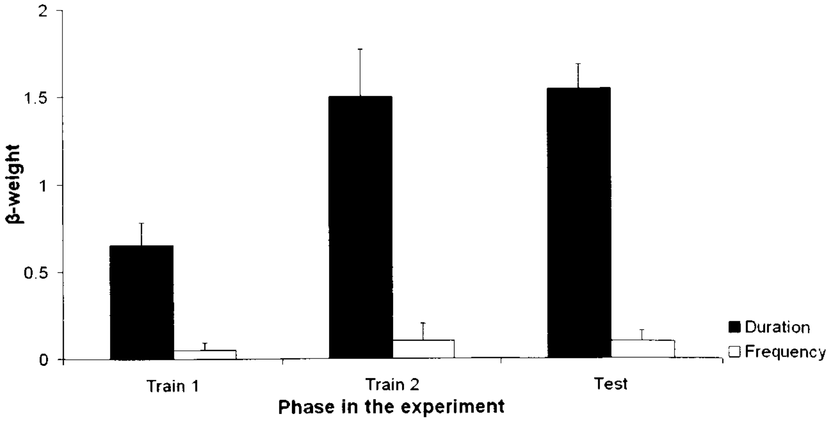

Fig. 3. β-weights for formant frequency and duration in the three phases of the nonspeech experiment, condition 1: relevant variation in one dimension (duration).

Table 2
Logistic regression results for the nonspeech experiment for each condition

	Condition 1 (1 relevant dimension, $N = 12$)				Condition 2 (2 relevant dimensions, $N = 12$)			
				Training part 1				
	Mean β	σ	Uni	Multi	Mean β	σ	Uni	Multi
Duration	0.65	0.44	10	0	0.08	0.05	3	4
Frequency	0.05	0.04	0		0.11	0.09	0	
				Training part 2				
	Mean β	σ	Uni	Multi	Mean β	σ	Uni	Multi
Duration	1.50	0.94	11	0	0.10	0.08	1	6
Frequency	0.10	0.10	0		0.21	0.17	1	
				Test				
	Mean β	σ	Uni	Multi	Mean β	σ	Uni	Multi
Duration	1.54	0. 47	12	0	0.70	0.32	8	1
Frequency	0.10	0.06	0		0.19	0.19	0	

Notes: β-weights are shown for both dimensions and the number of subjects using one (Uni) or both (Multi) dimensions significantly.

Condition 1 presents a clear picture of learning of a category structure with one relevant dimension of variation. Subjects discover the relevant dimension early in the training phase, but their performance improves throughout their training, as evidenced by the increasing β-weights. The difference in the β-weight for duration between the first and second part of the training is significant (F (2,33) = 6.895, $p < 0.05$). Post hoc analyses of the β-weights with the Student–Newman–Keuls test show that Training Part 2 and the Test form a homogeneous subset ($p < 0.05$), while Training Part 1 differs from both Training Part 2 and Test. The difference in the β-weight for frequency between the first and second parts of the training is not significant (F(2,33) = 1.725, $p > 0.2$).

The results of condition 2 are harder to interpret. Figure 4 displays the β-weights for both dimensions, while Table 2 shows the values of the β-weights and the number of subjects preferring a unidimensional versus a multidimensional solution. As the size of the β-weights indicates, performance is not very good compared with condition 1. There are significant effects of training in condition 2, for frequency (F (2,33) = 13.223, $p < 0.05$) and duration (F (2,33) = 3.901, $p < 0.05$). The individual data show that in Training Part 2, 6 out of 12 subjects use both dimensions significantly in their categorization. This learning does not, however, generalize to the test phase, when only one subject still uses a multidimensional strategy.

Post hoc analyses of the β-weights for the dimension duration show that Training Part 1 and Training Part 2 are one homogeneous subset, as is the test phase. So duration is not used more in the second part of the training phase, compared to the first. In the test phase, the β-weight for duration is different from that in the training. Without the information provided by the supervision and the category structure, subjects start using duration in their categorization. For frequency, the situation is different; the post hoc analyses reveals that subjects learn to use frequency during the training phase; Training Part 1 and Training Part 2 are different (heterogeneous) subsets according to the Student– Newman–Keuls test. Training Part 1 and Test do not differ significantly, which indicates that the learning does not generalize to the test phase.

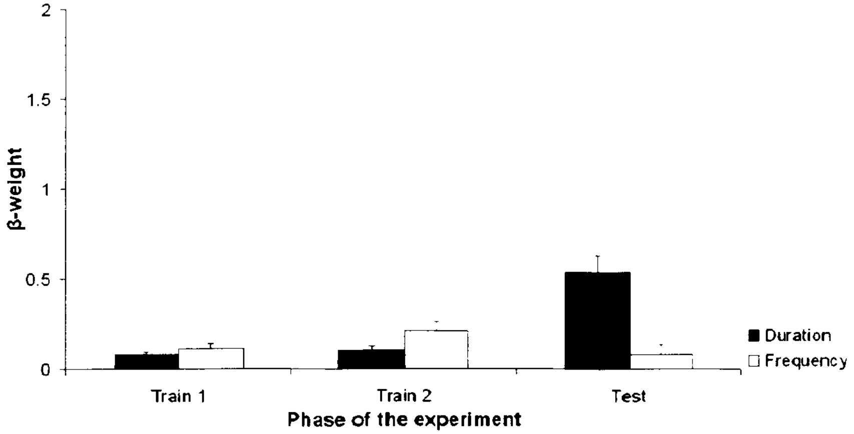

Fig. 4. β-weights for formant frequency and duration in the three phases of the nonspeech experiment, condition 2: relevant variation in both dimensions (formant frequency and duration).

Comparing condition 1 and condition 2 shows that real learning took place in condition 1, with performance improving during the training and reaching ceiling in the test. In condition 2, on the other hand, only about half of the subjects learned to use both dimensions in the training. In the test phase of condition 2, however, subjects almost invariably changed their categorization strategy, choosing one dimension to categorize the stimuli.

This study showed that, as expected on the basis of the evidence from visual category learning, participants who received feedback about their categorizations readily learned distributionally defined categories, and they were able to ignore irrelevant variation. However, even with feedback, performance in learning to use *two* distinct dimensions for separating categories defined over both dimensions was difficult. In the next section, the generality of this result was evaluated by testing listeners on synthesized tokens of nonnative vowel sounds.

4. Learning of speech categories

Our investigation of speech category learning was directly analogous to the nonspeech study. To assess possible pre-existing categorization tendencies for these speech stimuli, the experiment began with a pretest phase. Because the pretest stimuli and procedure were identical to the test phase, any difference between the pretest and the test phase can be attributed to the effect of the training phase.

The stimuli were synthesized versions of the Dutch front vowels /ø/, /y/, and /ʏ/ as in the Dutch words *feut* (/føt/, 'freshman'), *fuut* (/fyt/, 'grebe'), and *fut* /fʏt/, 'energy'). The differences between these vowels can be described by the duration and frequency of the first formants [Nierop, Pols, and Plomp (1973), Pols, Tromp, and Plomp (1973)]. The first formant of both /ø/ and /ʏ/ is approximately 441 Hz (about 8 ERB), while that of /y/ is approximately 305 Hz (about 10 ERB). As for duration, /ʏ/ and /y/ are short vowels of approximately 98 ms (about 46 D) while /ø/ is longer, at an average of 161 ms (about 51 D). All stimuli were generated using the PRAAT speech synthesis program [Boersma (2001)]. With the exception of frequency, all formants were kept constant for all stimuli (see Table 3). The values for the training stimuli were obtained by random sampling from two distributions. Careful listening by Dutch listeners (the first and second authors) revealed that the means of the categories still qualified as acceptable examples of the two Dutch vowels. In condition 1, the categories were defined by separation in the dimension of duration, with equivalent irrelevant variation in first formant frequency. In condition 2, the categories were defined by separation in both dimensions – duration and formant frequency.

Summary statistics for the stimuli are presented in Table 3. The test stimuli had the same design as in the nonspeech study: a grid with dimensions of equal psychological range.

All subjects were students from the University of Wisconsin, Madison. Some participated in exchange for course credit, while others were paid for their participation.

Table 3
Distributional characteristics of the stimuli for training condition 1 (relevant variation in one dimension), training condition 2 (relevant variation in two dimensions), and for the test phase of the speech experiment

	Training stimuli					
	Category A “/ø/”			Category B “/ʏ/”		
	Means	σ	ρ	Means	σ	ρ
Condition 1	50.7 DUR 10.0 ERB	0.5 DUR 0.2 ERB	−0.10	45.9 DUR 10.0 ERB	0.8 DUR 0.2 ERB	−0.05
Condition 2	49.7 DUR 10.1 ERB	2.6 DUR 0.2 ERB	−0.95	46.21 DUR 9.9 ERB	3.7 DUR 0.2 ERB	−0.95

	F2	F3	F4	F5
Fixed formants	19.6 ERB	22.3 ERB	26.2 ERB	28.2 ERB

	Test stimuli			
	Mean	Min	Max	Step size
Duration	48.5 DUR	45.85 DUR	50.8 DUR	0.5 DUR/step
F1	10 ERB	9.9 ERB	10.5 ERB	0.3 ERB/step

All were native speakers of English and did not speak another Germanic language or any other language with front–rounded vowels. Before the experiment, subjects filled out a consent form. In condition 1 there were 10 subjects, and in condition 2 there were 18 subjects.

Subjects were seated in a soundproof booth and were provided with a response box with four buttons and a light above each button. The experiment consisted of three phases; a pretest phase, a training phase, and a test phase. In the pretest and in the test phase, subjects listened to 196 (4 × 49) stimuli from the speech test continuum and categorized them as belonging to group A or B, without receiving feedback. In the training phase, subjects listened to 448 (2 × 2 × 112) speech training stimuli. The stimuli were presented in randomized blocks of 224. Training Parts 1 and 2 were analyzed separately. As in the nonspeech experiment, the difference between the conditions was in the number of relevant dimensions. In condition 1, the stimuli from a structure varied irrelevantly in one dimension (formant frequency) and relevantly in another (duration). In condition 2, the stimuli manifested relevant variation in both dimensions (see the left and middle panels of Figure 2). Subjects had to assign a stimulus to the leftmost or the rightmost of four buttons. A light above the button indicated which button was the correct response. If the light turned on, the categorization was correct; if not, the response was incorrect. The light stayed on for 400 ms, after which the next trial began.

Again, the results for condition 1 presented a clear picture (see Figure 5). There was a clear learning effect for duration (F (3, 36) = 12.365, $p < 0.05$). Subjects learned to use duration during the training phase, and retained this ability. There was also a small, but significant, effect for F1 (F (3, 36) = 6.6, $p < 0.05$) showing that subjects did not ignore this dimension entirely. Post hoc analyses with the Student–Newman–Keuls test

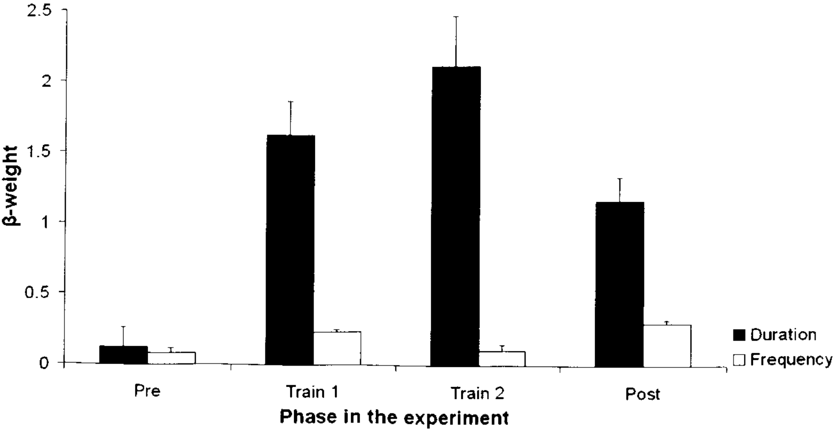

Fig. 5. *β*-weights for formant frequency and duration in the four phases of the speech experiment, condition 1: relevant variation in one dimension (duration).

revealed that for duration there were three separable homogeneous subsets, the pretest, the training, and the post-test ($p < 0.05$). For frequency, Training Part 1 was different from the other three parts ($p < 0.05$). Table 4 shows that in the pretest there was no preference for using either duration or frequency in categorizing the stimuli. After the first training phase, 9 of the 10 subjects consistently used duration and this generalized to the test phase.

As in experiment 1, the results of condition 2 are harder to interpret. According to Figure 6, the average use of both conditions increases. However, because the increase is small, this effect is not significant for duration ($F\,(3, 68) = 0.189, p < 0.904$) or for frequency (though the latter effect is stronger: $F\,(3, 68) = 2.189, p < 0.097$). In the test phase the learning effect washes away, and subjects hardly use duration anymore. The post hoc tests support this interpretation, as all phases of the experiment are in the same subset.

The individual subjects' data in Table 4 show the same pattern: although subjects exhibited a preference for frequency in the pretest, this changed in the first phase of the training, where both duration and frequency were used by five subjects. Six subjects already used both dimensions simultaneously, and this increased to nine subjects in the second phase of the training. This learning effect did not generalize to the test phase, where subjects preferred to use frequency in categorizing the stimuli.

Comparing conditions 1 and 2 yields the same result as in experiment 1. In condition 1, subjects learned to use the relevant dimension, whereas in condition 2 using both dimensions was considerably more difficult. Still, half of the subjects learned to use two dimensions in the second phase of the training, and one-third of the subjects used two dimensions in the test phase.

Table 4
Logistic regression results for the speech experiment for each condition

	Condition 1 (1 relevant dimension, $N = 10$)				Condition 2 (2 relevant dimensions, $N = 18$)			
				Pretest				
	Mean β	σ	Uni	Multi	Mean β	σ	Uni	Multi
Duration	0.30	0.14	3	0	0.33	0.10	2	4
Frequency	0.13	0.04	4		0.86	0.19	10	
				Training part 1				
	Mean β	σ	Uni	Multi	Mean β	σ	Uni	Multi
Duration	1.62	0.24	10	0	0.44	0.18	5	6
Frequency	0.11	0.02	0		0.96	0.08	5	
				Training Part 2				
	Mean β	σ	Uni	Multi	Mean β	σ	Uni	Multi
Duration	2.11	0.36	9	0	0.51	0.12	2	9
Frequency	0.23	0.04	0		1.06	0.21	4	
				Test				
	Mean β	σ	Uni	Multi	Mean β	σ	Uni	Multi
Duration	1.16	0.17	9	0	0.20	0.06	0	6
Frequency	0.08	0.02	1		0.99	0.19	10	

Notes: β-weights are shown for both dimensions and the number of subjects using one (Uni) or both (Multi) dimensions significantly.

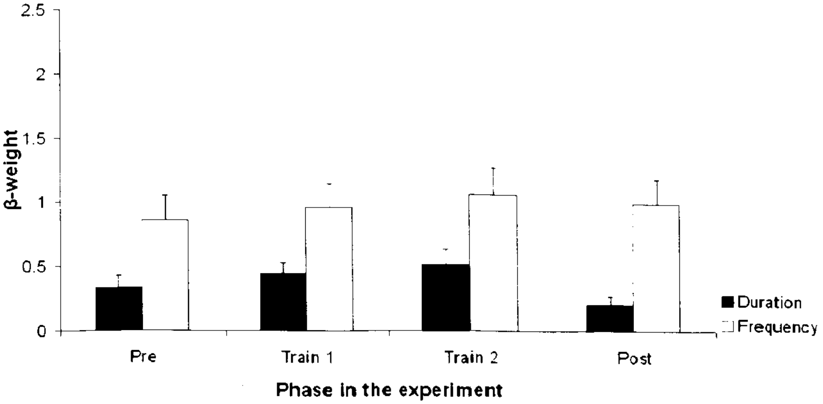

Fig. 6. β-weights for formant frequency and duration in the four phases of the speech experiment, condition 2: relevant variation in both dimensions (formant frequency and duration).

5. Conclusion

Auditory category learning for both speech and nonspeech stimuli is strikingly difficult. Our data show that listeners can relatively quickly learn a unidimensional categorization, but a categorization involving the simultaneous use of two dimensions is very hard to acquire. The unidimensional categorization was robustly acquired even though listeners had to ignore irrelevant variation in another dimension. The multidimensional categorization remained hard even though listeners had at their disposal two sources of information about the category structure, namely both the distributional characteristics of the category exemplars and feedback regarding their category judgments.

Learning of a multidimensional category structure is, we argue, the task facing infants who are learning to map the phonemes of their native language onto their psychophysical space. Yet we have shown that this task is extremely difficult for adult listeners. There are several possible explanations for this discrepancy.

First, the differences between the learning capacities of infants and adults may be greater than we had hypothesized. Adults may simply have lost the pattern recognition skills available to infants. Also, adults have already acquired a phonemic inventory, which may interfere with learning new phonetic categories or analogous auditory categories [Best and Strange (1992)]. Second, it is conceivable that explicit feedback leads subjects to attempt to formulate a verbal description or rule for use in their categorizations. Such verbal rules generally cannot separate multidimensionally varying categories, as argued by Ashby and colleagues [Ashby et al. (1998), Maddox, Ashby, and Waldron (2002), Maddox, Bohil, and Ing (2004)]. Third, infants receive a lot more exposure in acquiring their phonetic categories. Perhaps auditory category learning, unlike visual category learning, requires a very large number of exposures.

All current proposals for how infants spontaneously learn phonetic categories are distributional learning accounts in which infants are argued to perform statistical clustering over large numbers of isolated tokens of speech sounds. Experimental evidence with infants is compatible with this notion in broad outline, but in fact surprisingly little is known about the learning of auditory categories, either in infancy or in adulthood. The present experiments used techniques borrowed from related studies in the visual modality, presenting subjects with extensive exposure to distributionally defined categories with dimensions of variation known to be discriminable. Even with stimuli varying along two dimensions over a range of 20 just noticeable differences, though, and with supervised training, category learning performance in adults for both speech and nonspeech stimuli was poor and inconsistent, and collapsed soon after the close of training.

From a methodological point of view, of course, it is encouraging that the results for the nonspeech stimuli and the speech stimuli were so similar. This supports our assumption that the nonspeech stimuli we used were good counterparts of the speech stimuli. The former were not recognized as being speech, yet they were analyzed in a similar manner. However, the subjects' overall poor performance in auditory category acquisition remains puzzling. What learning occurred in our study seemed, moreover, rather fragile. In the test phase of our experiments, we observed in all conditions, and for both nonspeech and speech, that subjects "unlearned" the categorization rule they had previously mastered. We suggested above that this resulted from the stimulus configuration with which listeners were presented in the test phase. The use of a grid with uniformly spaced stimuli to assess the psychophysical space of a listener is, in fact, a well-known technique in the field of phonetics and phonology. The lack of information in the distribution of the stimuli is intended to prevent subjects from changing their categorization tendencies. However, this is not what happened; our listeners apparently picked up on the fact that in the test phase both dimensions were equally relevant, and altered their categorizations to suit.

Studies of auditory perceptual learning with respect to already known phonetic categories have shown that adult listeners exhibit remarkable flexibility in adjusting the boundaries of the phoneme categories of their native language [Norris, McQueen, and Cutler (2003), Evans and Iverson (2004), Eisner and McQueen (2005)]. Such adjustments enable listeners to adapt quickly to new speakers and new dialects; eventually, perceptual adjustments of this nature also underlie pronunciation changes across whole language communities. The listeners in our experiments seemed to maintain analogous flexibility towards the use of auditory information in the input; where they performed poorly was in converting the information into confident categorical decisions. The categories which were most speech-like, in that they were defined by multidimensional variation, were the hardest for these adult listeners to acquire. By contrast, the unidimensionally defined categories were reasonably well learned despite substantial irrelevant variation in a second dimension. These results thus point to significant gaps in our current understanding of auditory category learning, and of how infants acquire phonetic categories. The next step is, clearly, to endeavor to close these gaps.

References

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

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., L.A. Alfonso-Reese, A.U. Turken and E. Waldron (1998), "A neuropsychological theory of multiple systems in category learning", Psychological Review 105:442–481.

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

Aslin, R.N., P.W. Jusczyk and D.B. Pisoni (1998), "Speech and auditory processing during infancy: Constraints on and precursors to language", in: D. Kuhn and R. Siegler, eds., Handbook of Child Psychology, 5th edition, vol. 2: Cognition, Perception and Language. W. Damon, series editor (Wiley, New York) 147–198.

Behnke, K. (1998), "The acquisition of phonetic categories in young infants: a self organising neural network approach", Doctoral Dissertation (University of Nijmegen), MPI Series in Psycholinguistics, 13.

Best, C., G.W. McRoberts and N.M. Sithole (1988), "Examination of the perceptual reorganization for speech contrasts", Journal of Experimental Psychology: Human Perception and Performance 14:245–260.

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

Boersma, P. (2001), "PRAAT, a system for doing phonetics by computer". Glot International 5:341–345.

Cameron Marean, G., L.A. Werner and P. Kuhl (1995), "Vowel categorization by very young infants", Developmental Psychology 28:163–405.

Cutler, A. and M. Broersma (2005), "Phonetic precision in listening", in: W. Hardcastle and J. Beck, eds., A Figure of Speech (Lawrence Erlbaum, Mahwah, NJ), 63–91.

Eisner, F., and J.M. McQueen (2005), "The specificity of perceptual learning in speech processing", Perception & Psychophysics 67:224–238.

Evans, B. G., and P. Iverson (2004), "Vowel normalization for accent: an investigation of best exemplar locations in northern and southern British English sentences", Journal of the Acoustical Society of America 115:352–361.

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

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

Hillenbrand, J. (1983), "Perceptual organization of speech sounds by infants", Journal of Speech and Hearing Research 26:268–282.

Jusczyk, P.W. (1997), The Discovery of Spoken Language (MIT Press, Cambridge, MA).

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

Kornai, A. (1998), "Analytic models in phonology", in: J. Durand and B. Laks, eds., The Organization of Phonology: Constraints, Levels and Representations (Oxford University Press, Oxford) 395–418.

Kuhl, P.K. (1985), "Categorization of speech by infants", in: J. Mehler and R. Fox, eds., Neonate Cognition: Beyond the Blooming Buzzing Confusion (Lawrence Erlbaum Associates, Hillsdale, NJ) 231–263.

Kuhl, P.S. (2000), "A new view of language acquisition", Proceedings of the National Academy of Sciences of the United States of America 97:11850–11857.

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

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

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

Maye, J., and L. Gerken (2000), "Learning phonemes without minimal pairs", in: S.C. Howell, S.A. Fish, and T. Keith-Lucas, eds., Proceedings of the 24th Annual Boston University Conference on Language Development, vol. 2, (Cascadilla Press, Somerville, MA) 522–533.

Maye, J. and L. Gerken (2001), "Learning phonemes: How far can the input take us?", Paper presented at the Boston University Conference on Language Development (Somerville, MA).

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

Minda, J.P. and J.D. Smith (2001), "Prototypes in category learning: the effect of category size, category structure, and stimulus complexity", Journal of Experimental Psychology: Learning, Memory and Cognition 27:775–799.

Minda, J.P., and J.D. Smith (2002), "Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation", Journal of Experimental Psychology: Learning, Memory and Cognition 28:275–292.

Nierop, D.J.P.J.V., L.C.W. Pols and R. Plomp (1973), "Frequency analysis of Dutch vowels from 25 female speakers", Acustica 29:110–118.

Norris, D., J.M. McQueen and A. Cutler (2003), "Perceptual learning in speech", Cognitive Psychology 47:204–238.

Nosofsky, R.M. (1992), "Exemplar-based approach to categorization, identification and recognition", in: F.G. Ashby, ed., Multidimensional Models of Perception and Cognition (Lawrence Erlbaum Associates, New York) 363–393.

Peterson, G.E., and H.L. Barney (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.

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

Pols, L.C.W., H.R.C. Tromp and R. Plomp (1973), "Frequency analysis of Dutch vowels from 50 male speakers", Journal of the Acoustical Society of America 53:1093–1101.

Saffran, J. (2001), "Words in a sea of sounds: the output of infant statistical learning", Cognition 81:149–169.

Saffran, J.R., R.N. Aslin and E.L. Newport (1996),"Statistical learning by 8-month-old infants", Science 274:1926–1928.

Smits, R., J. Sereno and A. Jongman (in press), "Categorization of sounds", Journal of Experimental Psychology: Human Perception and Performance.

Werker, J.F., and R.C. Tees (1984), "Cross-language speech perception: evidence for perceptual reorganization during the first year of life", Infant Behavior and Development 7:49–63.