Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 50 (2008) 109–125
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Supervised and unsupervised learning of multidimensionally varying non-native speech categories

Martijn Goudbeek [a,*], Anne Cutler [a,b], Roel Smits [a]

[a] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[b] MARCS Auditory Laboratories, University of Western Sydney, Australia

Received 24 October 2006; received in revised form 24 July 2007; accepted 25 July 2007

Abstract

The acquisition of novel phonetic categories is hypothesized to be affected by the distributional properties of the input, the relation of the new categories to the native phonology, and the availability of supervision (feedback). These factors were examined in four experiments in which listeners were presented with novel categories based on vowels of Dutch. Distribution was varied such that the categorization depended on the single dimension duration, the single dimension frequency, or both dimensions at once. Listeners were clearly sensitive to the distributional information, but unidimensional contrasts proved easier to learn than multidimensional. The native phonology was varied by comparing Spanish versus American English listeners. Spanish listeners found categorization by frequency easier than categorization by duration, but this was not true of American listeners, whose native vowel system makes more use of duration-based distinctions. Finally, feedback was either available or not; this comparison showed supervised learning to be significantly superior to unsupervised learning.

Keywords: Category learning; Non-native categorization; Statistical learning; Vowels; Duration; Frequency

1. Introduction

It can be extremely difficult for adults to learn a non-native phonetic distinction, especially to native or near-native levels of discrimination (Burnham et al., 1991; see Strange, 1995, for reviews). One important reason for this difficulty is that a native phonology is already available to determine how speech sounds should be categorized, and this interferes with the learning of new categorizations (Cutler and Broersma, 2005; Best and Tyler, 2007). The learning of non-native phonetic contrasts has been intensively studied, and has prompted the development of a number of theoretical accounts, such as the Speech Learning Model (SLM; Flege, 1995) and the Perceptual Assimilation Model (PAM; Best, 1995; Best et al., 1988). These models postulate several ways in which non-native speech sounds may or may not map to native categories: they may be categorized within the phonological system of the native language, be left uncategorized but still perceived as speech or, more rarely, be left unassimilated and thus not treated as speech at all.

In the latter case, as indeed predicted by PAM, category discrimination is good to excellent. For example, American listeners hear Zulu clicks as non-speech, but discriminate them as well as native Zulu listeners do (Best et al., 1988). Discrimination in the case where the non-native sounds are categorized in the native system depends on whether the new sounds map to the same or to separate native categories. The situation in which non-native speech sounds are considered as speech but left uncategorized (i.e., not mapped to native speech categories) also allows for a range of discrimination possibilities. This situation arises when no native phonetic categories are sufficiently close

* Corresponding author. Address: Department of Psychology, University of Geneva, 40, Boulevard du Pont d'Arve, 1205 Geneva, Switzerland. Tel.: +41 22 379 9207.
E-mail address: goudbeek@pse.unige.ch (M. Goudbeek).

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.07.003

to the non-native ones in phonetic space to make mapping possible. According to PAM, either one non-native category or both could be left uncategorized (Best, 1994, 1995; Best and Tyler, 2007). When only one category is left uncategorized, and the other mapped to a native category, discrimination can be very good. When both non-native categories are left uncategorized, discrimination can be poor or fairly good, depending on the distance of each non-native category to the closest native phoneme categories. SLM makes no clear predictions about discrimination, but it can be reasonably inferred that discrimination success in SLM hinges on the establishment of a new non-native category and the perceptual distance of this category to already established categories (Flege, 1995).

As can be inferred from the previous paragraphs, the mapping of native and non-native categories is usually tested in a discrimination paradigm, although discrimination and categorization are only indirectly related. The ability to discriminate speech sounds is a necessary condition for the ability to categorize speech sounds, but not a sufficient one. However, a speech recognition system with an excess of discrimination abilities without the accompanying categorization abilities is thought to be unlikely.

The influence of native categories on the perception of non-native categories has usually been studied either in naïve listeners with no knowledge of the non-native language in question, or in second-language learners who are attempting to acquire phoneme categories in the course of acquiring the language – its words and its structures (Best et al., 1988; Logan et al., 1991). The acquisition of a category distinction has however hardly ever been examined as a process in its own right, independently of the language acquisition process. An exception is the work of Francis and Nusbaum (2002) who examined the way in which native English listeners learned to attend to dimensions relevant in discriminating Korean stop consonants. Their results showed that in learning new categories, listeners restructure their perceptual space, at least for the duration of the experiment; they do this mainly by reweighting its existing dimensions (see Francis et al., 2000), but also to a certain extent by attending to dimensions that were previously unattended.

In the current study we attempted to focus on the factors which control the success of non-native category learning. We presented listeners with a vowel category distinction of an unfamiliar type, and examined their success in learning and being able to apply it, as a function of three experimental parameters: (1) the availability of supervision (feedback) during the learning process; (2) the number of dimensions with relevant or irrelevant variation; (3) the number and placement in phonetic space of native phoneme categories. For the latter comparison we made use of listeners with different native languages, varying in vowel repertoire. Our experiments used classic category learning procedures borrowed from studies of visual perception.

Whatever is being learned, the learning process will essentially depend upon the input that the learner receives. In visual category learning, the effects of the distributional properties of the input have been extensively studied (Ashby and Maddox, 1993; Nosofsky, 1990). Perceptual categories are defined, in this literature, as either points, collections of points, or distributions in a psychophysical space with continuous dimensions. Clearly, phonetic categories can be considered in the same way, and auditory category learning as equivalent to recognizing the statistical patterns in auditory input (see, e.g., Pierrehumbert, 2003, for such a proposal). When a listener hears a sound, this sound can be evaluated on a number of dimensions (e.g., duration, frequency) and mapped onto a point corresponding to its values in multidimensional auditory space. Sounds originating from the same category will be consistently mapped to the same area, and repeated exposure to categories leads to the formation of distinct "clouds" that listeners can start to associate with a category label.

The distributional structure of the input will crucially affect the way categorization decisions are made (Maye and Gerken, 2000, 2001). As an illustration, we generated four possible stimulus structures, as displayed in Fig. 1. Exposure to the stimulus structure in the upper left panel of Fig. 1 should encourage listeners to categorize using only dimension 1 and ignore dimension 2; this enables them to optimally separate the different categories as depicted by the crosses and the circles. Exposure to the stimulus structure in the lower left panel, in contrast, should encourage listeners to categorize using only dimension 2 and ignore dimension 1. That is, in the upper left panel, dimension 1 displays relevant variation and dimension 2 displays irrelevant variation, in the lower left panel this situation is reversed. Exposure to the structures on the right hand column should encourage listeners to use both dimensions in their categorization. A categorization strategy that uses only one dimension in categorizing the stimuli in the panels of the right hand column would lead to many incorrect decisions.

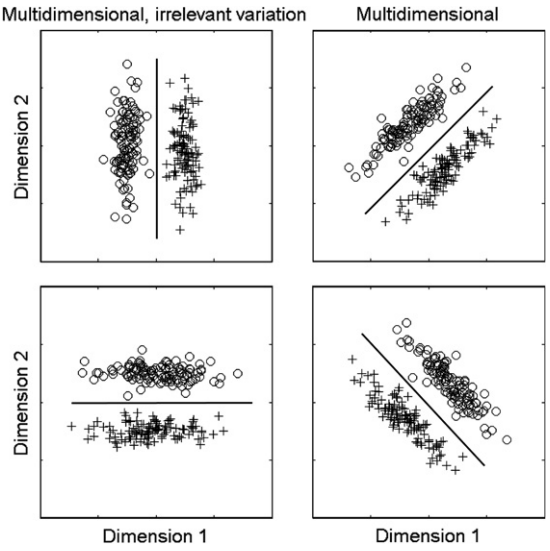

Fig. 1. Four possible category structures in a two-dimensional perceptual space. Lines represent the optimal solution to the categorization problem, different symbols represent separate category structures.

Visual category learning experiments have shown that subjects initially opt for a solution involving only one dimension (Feldman, 2000) and that they need the help of trial-by-trial feedback on the correctness of their response to start using more than one dimension in their categorizations (Ashby et al., 1998). In contrast to the categories involved in word recognition and language comprehension, the categories used in these experiments are arbitrary ones. Ashby et al. (1998) argue that there are two category learning systems, a verbal learning system and a procedurally based or implicit learning system. Initially, the verbal system has priority and tries to categorize the stimuli according to a relatively simple, verbalizable, rule involving only one dimension (e.g., high frequency sounds in category A, low frequency sounds in category B). Rules that are more complex and more difficult to verbalize such as "all short and high frequency sounds in category A" only enter the verbal system after all unidimensional options have been tried. The other, implicit, system is based on the learning of actual skills or procedures (in the present case, for categorization). This system does not have such a preference for unidimensional rules, but it learns much more slowly.

Studies of unsupervised learning of visual categories have shown that trial-by-trial feedback is not always necessary, but that there are characteristic limits to performance in unsupervised learning. Ashby et al. (1999) showed that participants confronted with a multidimensional categorization problem initially opt for a unidimensional solution (using only one dimension of variation in their categorizations). Their subjects had to categorize lines differing in length and orientation without the aid of supervision. Two groups of subjects encountered categories that were separable using only length or only orientation and where the other two dimensions displayed irrelevant variation. For the other two groups both dimensions were relevant (as in the right column of Fig. 1). By the end of the experiment, observers in the unidimensional conditions responded almost perfectly, whereas those in the multidimensional conditions were still not able to use both stimulus dimensions. Only in a follow-up experiment, in which trial-by-trial feedback was present, could subjects entertain a solution that used more than one dimension in their categorization. Homa and Cultice (1984) and Love (2002) also found limitations on unsupervised learning performance with complex problems. Homa and Cultice (1984) created connected dot patterns that differed in their level of distortion. Observers categorized these patterns with and without feedback. While feedback provided little benefit in learning low-distortion patterns, learning highly distorted patterns was only possible in the presence of feedback. Love (2002) investigated unsupervised learning with the category learning problems constructed by Shepard et al. (1961). Performance was best (73% correct) when only one dimension was relevant; with two relevant dimensions, accuracy dropped to 56% correct (Love, 2002).

Our experiments, modeled on these studies of visual category learning, all consisted of a pretest, a learning phase and a maintenance phase. The first panel of Fig. 2 shows the distributional structure of the pretest. The dimensions duration and formant frequency define the difference between the stimulus categories as described in Section 2.1.2. The stimuli are drawn from an equidistant grid with an equal range of variation as defined by just noticeable differences in both stimulus dimensions. In the pretest, this grid is intended to neutrally scan the listener's initial categorization tendencies.

The second through fourth panel show the learning phases of the various experiments. The second and third panel depict category structures called "unidimensional learning". In these cases, one dimension of variation is relevant to the classification of the vowel stimuli while another dimension is irrelevant for this classification. For optimal performance, listeners have to learn to use only the relevant dimension in their categorizations and ignore the other dimension. In the second panel, listeners are trained to use duration as a relevant dimension and ignore formant frequency variation, while in the third panel, listeners have to learn to use formant frequency in categorization but ignore duration. In contrast, in the type of category structure in panel four ("multidimensional learning"), both dimensions exhibit relevant variation. For optimal performance here, listeners have to learn to use both dimensions in categorization. The use of only one dimension would lead to a high proportion of incorrect categorizations. The learning phase of each experiment was analyzed in two parts (learning phases 1 and 2) to examine categorization behavior over time. All experiments were run in a single session with a short intermediate pause

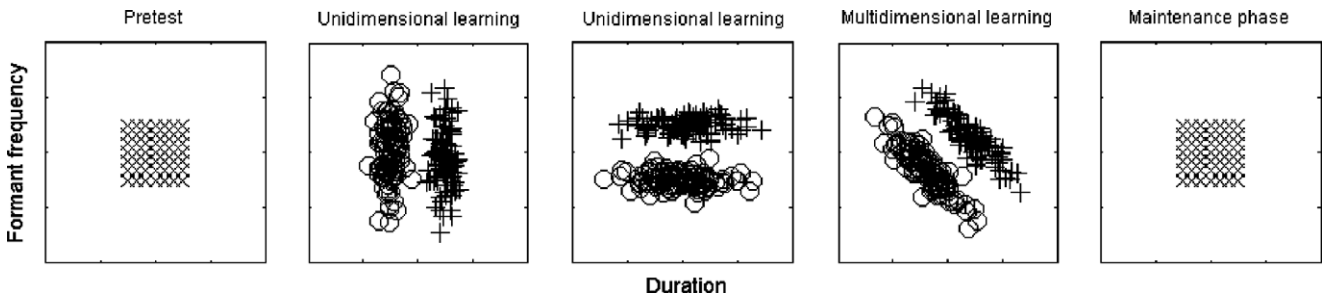

Fig. 2. The basic experimental design of Experiments 1–4: a pretest phase without distributional information, training phases with distributional information (either one or two relevant dimensions) and a maintenance phase that is identical to the pretest.

between the two phases, and all ended with a maintenance phase consisting of the same stimuli as the pretest. This again was intended to assess the listeners' perceptual space in the absence of distributional information. If listeners learned a new category structure in the learning phase and were able to transfer this learning to the maintenance phase, then performance in the maintenance phase should resemble that of the learning phase, and should differ from performance in the pretest.

The auditory categories which we manipulated were three Dutch vowels. The Dutch vowel inventory contains thirteen monophthongs and three diphthongs (Gussenhoven, 1999, in press; Booij, 1995), with three mid-to-high front-central vowels: /ʏ/ (as in /fʏt/, "fut"; "energy"), /y/ (as in /fyt/,"fuut"; "grebe") and /ø/ (as in /føt/, "feut"; "freshman"). These vowels differ primarily in the frequency of their first formant and in their duration. The sounds /ʏ/ and /y/ do not differ greatly in length, but /y/ has a lower first formant frequency, while the sounds /ø/ and /ʏ/ have similar frequency spectra, but /ø/ has a longer duration.

To examine the role of native category structure in category acquisition, we chose listeners whose native language has none of these vowels (not difficult, given that Dutch is cross-linguistically atypical in having so many vowels in this area of the vowel space). We sought to achieve the situation in which category assimilation does not happen, but sounds are still recognized as speech. This should happen when the sounds are located in a relatively empty area of phonetic space. A specific prediction about such cases is made by SLM: the distance of each non-native category to the closest native phoneme categories will affect category acquisition success, with more distinct categories being easier to learn (Flege, 1995; and see Aoyama et al., 2004, for confirmation of this prediction). Furthermore, the relevance of the required categorization dimensions to native categorization decisions will also play a role. We tested listeners with two language backgrounds: Castilian Spanish and American English.

Spanish has a relatively small vowel inventory of five vowels: /i/, /e/, /a/, /o/, and /u/ that differ in height, backness and roundedness (Hammond, 2001; Bradlow, 1995; Flege, 1989). These articulatory dimensions correlate with the first and second formants of the acoustic signal. The high vowels /i/ and /u/ have low values for F1, whereas the higher values of F1 are associated with the mid (/o/ and /e/) and low (/a/) vowels. Backness and roundedness are associated with low values for F2 (/u/, /o/, and /a/) whereas front and unrounded vowels (/e/ and /i/) have a high value for F2 (Bradlow, 1995). Importantly, Spanish does not have durational contrasts between vowels as Dutch does in contrasting /ʏ/ and /ø/ (Booij, 1995; Gussenhoven, in press). Furthermore, the Spanish vowels are all located at the periphery of the F1/F2 vowel space. Thus the Dutch vowels closer to the center of F1/F2 space occupy an empty part of Spanish vowel space, and are arguably too far from any native Spanish vowel category for assimilation to be possible.

American English has, like Dutch, a large vowel inventory: 15 vowels (Ladefoged, 1999), including the central vowel schwa. Nonetheless, the area in vowel space that corresponds to the three Dutch vowels /ʏ/, /y/, and /ø/ is for the most part unused in American English. Thus again, the vowels should not be assimilated to a native category. However, given that American English has more vowel categories and does not restrict vowels to the periphery of F1/F2 space, listeners may be more inclined to integrate the closer exemplars of the Dutch stimuli into their phonological system. Furthermore, American English vowels exhibit significant variation in duration, the smallest minimum-to-maximum range in the adult data of Hillenbrand et al. (1995) being 182 ms for /I/, and the largest 272 ms for /u/. English listeners are highly sensitive to vowel duration as a cue to postvocalic consonantal voicing (Lisker, 1978), and cannot refrain from using this cue even when it is uninformative (Broersma, 2006). They have also been shown to be sensitive to duration when categorizing vowels (McAllister et al., 2002; Nearey, 1989). For all these reasons, we predict that acquisition of these Dutch vowel categories should be easier for American English than for Spanish listeners.

In Experiment 1, we compared supervised learning by Spanish listeners of the contrast between /ø/ (longer duration) and /ʏ/ (shorter duration), versus the contrast of /ʏ/ (higher F1) with /y/ (lower F1). In Experiment 2, unsupervised learning of the same contrasts was investigated. In Experiment 3, supervised learning of the duration-based distinction between /ø/ and /ʏ/ by American English listeners was addressed, for comparison with the Spanish listeners'performance on the same task. Finally, in Experiment 4 American English listeners were trained, with supervision, on the multidimensional distinction between /ø/ (longer duration and high F1) and /y/ (shorter duration and lower F1). To categorize these stimuli successfully,listeners must use both dimensions at once, a task that is difficult for listeners of various language groups (Flege and Hillenbrand, 1986).

2. Experiment 1

Experiment 1 investigated supervised learning of category distinctions based on one relevant dimension (either duration or formant frequency) in Spanish listeners.

2.1. Method

2.1.1. Subjects

Twenty Spanish exchange students from the Radboud University Nijmegen participated in the experiment (ten in each condition). None spoke another language besides English, but most were engaged in learning Dutch. They rated their proficiency in English on a five point scale (0 = bad, 3 = average, 5 = good) as above average ($\mu = 3.5$, $\sigma = 0.90$, $N = 20$). Their proficiency in Dutch was low, with only a few subjects rating proficiency as

being present at all ($\mu = 1.4$, $\sigma = 0.89$, $N = 4$). All listeners reported normal hearing and were within the normal undergraduate age range. After the experiment they filled in a questionnaire intended to assess whether they recognized the stimuli as vowels; they all qualified the stimuli as such.

2.1.2. Stimuli

The categories of both conditions each had one relevant dimension of variation (see the second and third panel of Fig. 2). We defined the two categories as probability density functions in a multidimensional formant frequency × duration space. The nature of the probability density functions (their means and covariance matrices) governed the relevance of each dimension for category judgments. Fig. 1 displays the actual stimulus distributions used in the experiments.

In Condition 1, the variation in duration was relevant, whereas formant frequency varied irrelevantly with respect to category membership. The means of the two categories corresponded to the Dutch vowels /ʏ/ and /ø/ as in the Dutch words "fut" (/fʏt/, 388 Hz and 120 ms) and "feut" (/føt/, 392 Hz and 162 ms). These vowels differ from each other primarily in the duration dimension with /ø/ being a lengthened version of /ʏ/ (Booij, 1995; see also Flege, 1992 for further discussion concerning the relationship between English and Dutch vowels). Native Dutch listeners respond to gated fragments of /ø/ with a predominant response of /ʏ/ (Smits et al., 2003). In Condition 2, duration varied irrelevantly and formant frequency was systematically varied. The means of the two categories corresponded to the Dutch vowels /ʏ/ and /y/ as in the Dutch words "fut" (/fʏt/, 388 Hz, 102 ms) and "fuut" (/fyt/, 328 Hz 113 ms). These vowels differ from each other primarily in the frequency of their first formant (formant frequency) with /y/ being a higher (more fronted) version of /ʏ/ (Booij, 1995). Native listeners do not confuse gated fragments of these vowels (Smits et al., 2003). Both vowels occur commonly in Dutch. The vowels were synthesized using the PRAAT speech synthesis program (Boersma and Weenink, 2003).

Careful listening by native Dutch listeners (the first and third author) confirmed that the means of the categories qualified as good examples of the two Dutch vowels. The values for the learning stimuli were obtained by random sampling from the two stimulus distributions.

The pretest and maintenance stimuli were identical in both conditions. The stimulus values for the pretest and the maintenance phase were obtained from an equidistantly spaced grid with duration and formant frequency as dimensions (see the rightmost panel of Fig. 2). The formant frequency values in the grid ranged between the means of the stimuli from the learning phase. The range of stimulus duration expressed in just noticeable differences (jnds) was equal

Table 1
Stimulus characteristics of the phonetic categories used in Experiments 1–3

	Learning stimuli					
	Category A "/ø/" as in *feut*			Category B "/ʏ/" as in *fut*		
	Means	σ	ρ	Means	σ	ρ
Condition 1 (duration relevant)	**52.2 D** **165 ms** 9.1 ERB 392 Hz	0.34 D 12.4 ms 1.88 ERB 127.0 Hz	−0.10	**50.1 D** **102 ms** 9.1 ERB 388 Hz	0.28 D 6.6 ms 1.8 ERB 120 ms	−0.08
	Category A "/y/" as in *fuut*			Category B "/ʏ/"as in *fut*		
Condition 2 (frequency relevant)	50.4 D 113 ms **8.16 ERB** **328 Hz**	1.2 D 33 ms 1.3 ERB 87.7 Hz	−0.08	50.1 D 102 ms **9.1 ERB** **388 Hz**	0.28 D 6.6 ms 1.8 ERB 120 Hz	−0.10

	Maintenance stimuli			
	Mean	Min	Max	Stepsize
Duration	51.1 D 131 ms	50.0 D 101 ms	52.2D 166 ms	0.15 D/step 5.9 ms/step
Frequency	9.0 ERB 375 Hz	7.8 ERB 299 Hz	10.2 ERB 457 Hz	0.17 ERB/step 11.7 Hz/step
	F2	F3	F4	F5
Fixed formants	19.6 ERB 1657 Hz	22.3 ERB 2292 Hz	26.2 ERB 3607 Hz	28.2 ERB 4845 Hz

The rows presenting the stimuli of Conditions 1 (duration relevant) and 2 (formant frequency relevant) list stimulus duration in ms and D (the perceptual counterpart of ms) and the values for the first formant in Hz and ERB (the perceptual counterpart of Hz). Any deviation of ρ from 0 is due to sampling. Both conditions have the same maintenance phase stimuli. The mean, minimal, and maximal duration and formant frequency values of these are listed. Means for the dimensions that vary in each condition are in boldface. The last row presents the values of the four fixed formants F2–F5 used in the generation of all stimuli. Bandwidths were set at 10% of the frequency.

to the number of jnds of the frequency range. In the learning phase, the distance between the category means was 20 jnds; in the maintenance phase, the stimuli ranged between these means in seven equidistant steps.

Table 1 lists the summary statistics for the stimuli used in the pretest, the learning phase and the maintenance phase. Any differences between category A and B in formant frequency in Condition 1, or in duration in Condition 2, are entirely due to sampling variation.

2.1.3. Procedure

Listeners were seated in a soundproof booth in front of a computer screen and a two-button response box. Their task was to assign each stimulus to group A or B, using the two buttons. They were given no further information about the category labels or their response options.

The pretest (to detect preexisting categorization tendencies) and maintenance phase both consisted of 196 test stimuli (49 stimuli times 4 repetitions), whose values ranged between the mean values of both categories (see the "unidimensional learning" panels of Fig. 2). In the pretest and maintenance phase no feedback was given on listeners' categorizations. Once a participant had selected a category label on a trial, the monitor would display (the Spanish equivalent of) "next" for 700 ms and the next stimulus was played after a 200 ms delay. In the maintenance phase, listeners were asked to continue to categorize as they had done at the end of the learning phase.

The learning phase consisted of 448 stimuli (2 categories × 2 repetitions × 112 stimuli per category) presented at a comfortable level through Sennheiser headphones (HD 270). The stimuli were presented in random order in two sessions, separated by a brief rest period. All 112 stimuli from each category were presented once in each session. In contrast to the pretest and maintenance phase, trial-by-trial feedback *was* provided during the learning phase. Once participants had selected a category label on a trial, the monitor displayed (the Spanish equivalent of) "right" in green letters for 700 ms immediately following the response if the categorization was correct, and (the Spanish equivalent of) "wrong" in red letters if the categorization was incorrect. After the visual feedback disappeared, a 200 ms blank screen preceded the next stimulus.

After the experiment all participants filled out a questionnaire asking them whether they had recognized the sounds as speech, whether they had labeled the groups in any way, and whether they spoke a language besides English.

2.2. Results and discussion

2.2.1. Signal detection analysis

As a first analysis, percent correct and d' were calculated for the learning phases of each condition; these are displayed in Fig. 3a and b respectively. Recall that in the pretest and maintenance phase a stimulus grid was used without feedback, so correct and incorrect categorization

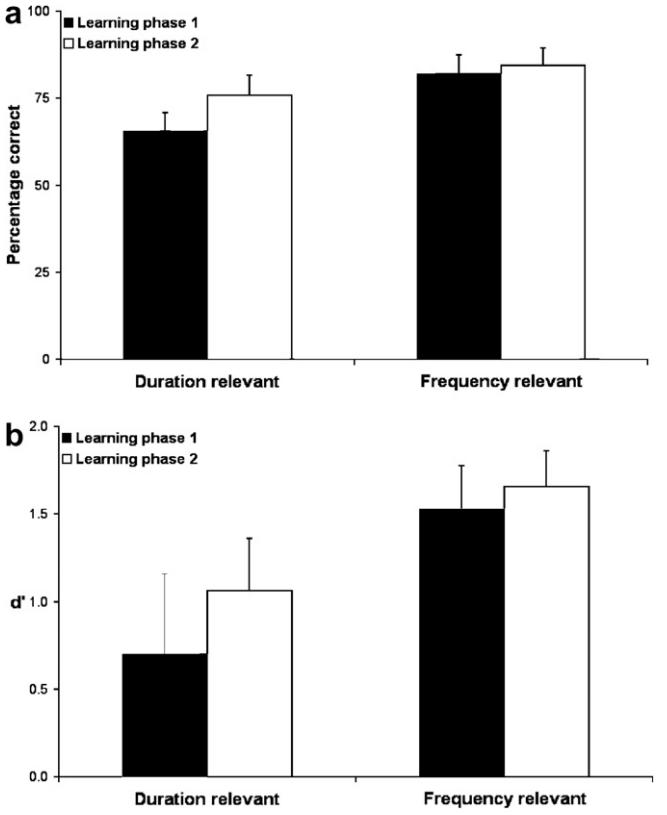

Fig. 3. Percentage correct (a) and d' (b) values for the two learning phases of Experiment 1, as a function of relevant dimension (duration versus frequency).

did not apply in these phases and hence these measures cannot be calculated; these phases are analyzed in detail in Section 2.2.2. Both percentage correct and d' show a clear increase in performance from the first to the second learning phase, indicative of a learning effect.

The percentage correct was significantly above chance in all phases of the experiment (minimum $t\,[9] = 2.92$, $p < 0.05$). To investigate the learning effect, we conducted an ANOVA with Learning phase (The first half versus the second half of the learning phase) as within-subjects variable and Condition (Duration relevant versus Frequency relevant) as between-subjects variable. This analysis showed the percentage correct to be significantly higher in the second learning phase ($F\,[1,18] = 6.30$, $p < 0.05$), which did not interact with Condition. Further, the analysis showed that when frequency was the relevant dimension subjects tended to achieve better categorization performance compared to when duration was the relevant dimension ($F\,[1,18] = 3.066$, $p < 0.097$).

In all learning phases and conditions, d' was significantly above zero (minimum $t\,[9] = 1.89$, $p < 0.05$), the value Macmillan and Creelman (1991) associate with identical distributions of perceptual effects of two stimuli. As with percentage correct, the main effect of Learning phase was significant ($F\,[1,18] = 7.58$, $p < 0.05$) and the effect of Condition was close to significance ($F\,[1,18] = 4.08$, $p < 0.06$). Again, the main effects did not interact.

The signal detection measures thus show a clear picture. There was a learning effect in both measures. There was no robust difference between conditions, although the condition in which frequency was the relevant dimension tended to be preferred. Because signal detection measures do not differentiate by dimension, and are not applicable to the pretest or maintenance phase, all four experimental phases were also analyzed with logistic regression.

2.2.2. Logistic regression

The binary choice design (every answer is either category A or category B) is very well suited to a logistic regression (Agresti, 1990). A logistic analysis yields two β-weights (which can be significant or not) indicating the extent to which each dimension explains the variation in the data. These β-weights are calculated for each listener individually and then averaged. To probe for learning, the two learning phases were analyzed separately. Table 2 shows mean β-weights and standard deviations of the dimensions duration and formant frequency for the pretest, the two learning phases, and the maintenance phase.

In addition to β-weights, a logistic regression procedure also gives a significance level, indicating whether a β-weight differs from zero and contributes significantly to the regression model. If the level was not significant for a given dimension, we concluded that listeners did not use this dimension in their categorization. The columns "Uni"and "Multi" in Table 2 show how many subjects made significant use of one or of both dimensions, respectively. These categories are mutually exclusive and the few subjects who used neither dimension have been omitted (given N, this number can be easily calculated).

Table 2
Logistic regression results of Experiment 1 in which Spanish listeners were trained with supervision to categorize stimuli with relevant variation in one dimension and irrelevant variation in the other dimension

	Condition 1, duration relevant ($N = 10$)				Condition 2, F1 relevant ($N = 10$)			
	$\mu(\beta)$	$\sigma(\beta)$	Uni	Multi	$\mu(\beta)$	$\sigma(\beta)$	Uni	Multi
Pretest								
Relevant	1.01	0.63	2	3	1.16	1.80	2	2
Irrelevant	1.55	1.97	2		0.93	0.88	4	
Learning phase 1								
Relevant	0.71	0.75	5	3	1.67	1.74	7	1
Irrelevant	0.17	0.1	1		0.24	0.17	0	
Learning phase 2								
Relevant	1.52	1.55	6	2	1.73	1.39	7	1
Irrelevant	0.26	0.26	2		0.32	0.31	0	
Maintenance phase								
Relevant	1.99	1.61	4	3	2.94	2.58	5	2
Irrelevant	1.06	1.10	2		0.35	0.46	1	

The table displays the results of the pretest, learning phases and maintenance phase of Condition 1 (duration relevant) and Condition 2 (formant frequency relevant). The mean β-weights and their standard deviations as well as the number of listeners using one ("Uni") or both ("Multi") dimensions significantly are shown. Listeners using no dimension significantly are not shown.

The results in Table 2 show the sensitivity of listeners to the information provided to them (trial-by-trial feedback and distributional information). In all phases except the pretest, the mean β-weights for the relevant dimension were higher than those for the irrelevant dimension. There were some differences between Conditions 1 and 2, possibly reflecting the preference for formant frequency as a relevant dimension also indicated in the signal detection analysis. First, the β-weight for the relevant dimension in Condition 1 was low in the first learning phase, suggesting an a priori reluctance to use this dimension. Similarly, in the maintenance phase ignoring irrelevant durational variation (Condition 2) appeared to be easier than ignoring irrelevant formant frequency variation (Condition 1).

These effects were evaluated with an ANOVA with Part of the experiment and Dimension (Relevant versus Irrelevant) as within-subjects variables and Condition as between-subject variable. The learning effect was present in the overall preference for the relevant over the irrelevant dimension ($F\,[1, 18] = 7.86$, $p < 0.05$) and in the increase in mean β-weight as the experiment progressed ($F\,[3, 54] = 9.096$, $p < 0.05$). There was no difference in performance between Conditions ($F\,[1, 18] = 0.17$, n.s.). The preference of our listeners for formant frequency in the Pretest resulted in a significant interaction between Part of the experiment and Dimension ($F\,[1, 54] = 7.45$, $p < 0.05$).

The results of Experiment 1 show that Spanish listeners were clearly able to learn a non-native category distinction characterized by relevant variation along one dimension and irrelevant variation along another when provided with trial-by-trial feedback. Independently of whether the relevant dimension was relatively unfamiliar (recall that duration does not play a significant role in the Spanish vowel system) or very familiar (formant frequency), our listeners were sensitive to the cues provided to them and could maintain the distinction they learned in the maintenance phase.

The trial-by-trial feedback provided in this experiment is not often available to the language learner. Infants acquiring a first language must rely exclusively on distributional information, and adults learning a second language also rely principally on distributional information, with feedback applied more to production than perceptual performance. In the native language, lexical information can be used to fine-tune existing categories (Norris et al., 2003), but such information cannot by definition create new categories. Therefore, in Experiment 2, unsupervised learning of the same speech categories as in Experiment 1 is investigated.

3. Experiment 2

Experiment 2 investigated unsupervised learning of category distinctions based on one relevant dimension (either duration or formant frequency) in Spanish listeners.

3.1. Method

3.1.1. Subjects

Fourteen Spanish exchange students from the Radboud University Nijmegen participated in the experiment (six in Condition 1 and eight in Condition 2). None spoke another language besides English, but most were engaged in learning Dutch. Their proficiency in Dutch was low. All subjects reported normal hearing. Again, all listeners judged the stimuli to be vowels or very vowel-like on the post-experiment questionnaire.

3.1.2. Stimuli and procedure

All stimuli were as in Experiment 1, and the procedure in the pretest and maintenance phase was also as in Experiment 1. In the learning phases, however, in contrast to the procedure of Experiment 1, no trial-by-trial feedback was provided. In all four phases of the experiment, the subject's task was to assign each stimulus to group A or B, using the two-key button box, after which the monitor would display (the Spanish equivalent of) "next" for 700 ms and the next stimulus was played after a 200 ms blank screen. The experiment lasted approximately 45 min.

3.2. Results and discussion

3.2.1. Signal detection analysis

The signal detection measures percent correct and d' are presented in Fig. 4. It is clear that performance in Condition 2 was better than in Condition 1. The results show little indication of the learning effect found in Experiment 1 in the difference between learning phases.

Before statistically testing these observations we first tested whether performance differed significantly from chance. The chance level for an experiment without feedback is less obvious than in an experiment with supervision. In order to calculate percent correct, each response must be labeled "right" or "wrong". In supervised learning, this is done a priori by the experimenter. In unsupervised learning, however, the experimenter has to infer the listener's mapping of stimulus and category. Some listeners will associate one category with label A and the other with label B, while others will use the reverse pattern.

For each listener, the category most associated with response A was defined as category A for subsequent analysis. As a consequence, subjects always perform at or above chance level. Therefore, the chance level is not simply at 50% correct but has to be adjusted. We calculated the expected value for chance level for 224 stimuli from a binomial distribution and the transformed percent correct, leading to an adjusted chance level of 52.66%.

In Condition 1, when duration was the relevant dimension, percentage correct did not differ from chance in the first ($t\,[5] = 1.65$, n.s.) or second learning phase ($t\,[5] = 1.47$, n.s.). However, in Condition 2, when formant frequency was the relevant dimension, both the percentage correct of the first learning phase ($t\,[7] = 8.23$, $p < 0.05$) and that of the second learning phase ($t\,[7] = 7.66$, $p < 0.05$) differed significantly from chance. This difference between the two conditions also appeared in the main effect for Condition in the ANOVA ($F\,[1,12] = 7.77$, $p < 0.05$) with Learning phase as within-subject variable. Performance did not improve over time, judging by the absence of a significant effect of Learning phase ($F\,[1,12] = 0.012$, n.s.).

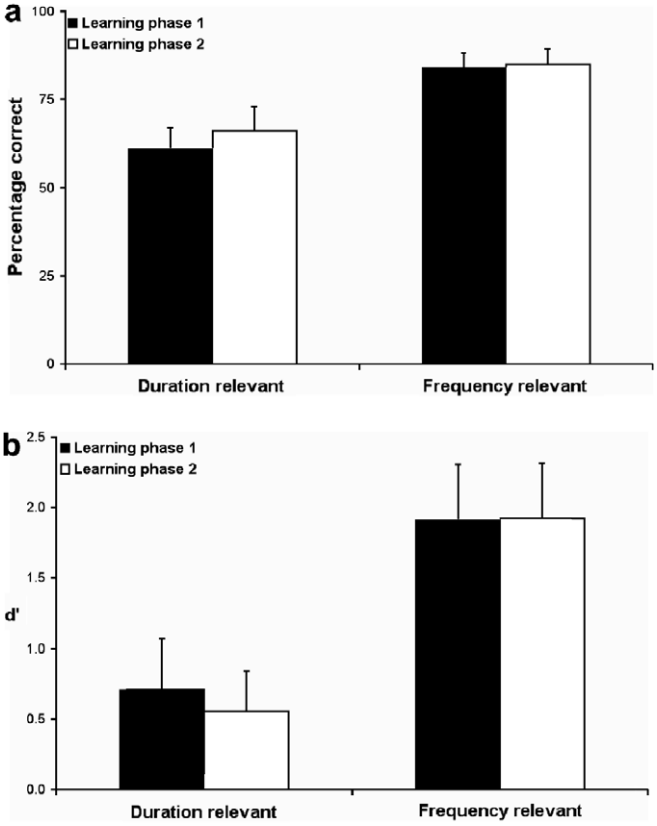

Fig. 4. Percentage correct (a) and d' (b) values for the two learning phases of Experiment 2, as a function of relevant dimension (duration versus frequency).

The d' results mirror those of the percentage correct. In condition 1 (duration relevant), none of the d's differed significantly from zero, whereas in condition 2 (formant frequency relevant) the d's of both learning phases differed significantly from zero ($t_{\min} = 4.83$). The d's in Condition 2 also were well above 1, the size traditionally associated with a true perceptible difference, so subjects were able to distinguish the two categories. In Condition 1, this was not the case. As with percentage correct, the two Conditions differed ($F\,[1,12] = 5.85$, $p < 0.05$) and there were no other significant effects.

Thus, according to the signal detection analysis, performance depended on which dimension was relevant. When formant frequency was the relevant dimension, listeners used this dimension appropriately; when duration was relevant, this was not the case.

3.2.2. Logistic regression

Table 3 shows the mean β-weights of Condition 1 (duration relevant) and Condition 2 (formant frequency relevant) for all four parts of the experiment.

Unsupervised learning of category structures with relevant variation in only one dimension appears to be difficult. An ANOVA with Dimension (Relevant versus Irrelevant), and Part of the experiment as within-subjects variables and Condition as between-subjects variable was conducted to assess any effect. For Dimension, no significant effect was found ($F\,[1,12] = 0.345$, n.s.). This means that participants did not show an overall preference for the relevant dimensions over the irrelevant one; instead they all preferred formant frequency over duration. While there was a significant effect of Part of the experiment ($F\,[3,36] = 21.04$, $p < 0.05$), this is probably due to the differences between the β-weights of the training phases and the pretest/maintenance phases of the different conditions and not to a real learning effect. Significant interactions between Part of the experiment and Condition ($F\,[3,36] = 7{,}25$, $p < 0.05$) and Part of the experiment and Dimension ($F\,[3,36] = 3{,}93$, $p < 0.05$) support this interpretation. To further investigate this, we conducted separate analyses per condition and for each combination of pretest/maintenance (Equidistant grid) phase and learning phase 1/learning phase 2 (Learning phase). This showed that the interactions were carried by the interaction between the Dimension and Equidistant grid (Pretest versus Maintenance phase) in Condition 2 ($F\,[1,7] = 7.928$, $p < 0.05$). So, only when formant frequency was relevant, listeners used the relevant dimension more in the maintenance phase than they used the irrelevant dimension in the pretest.

Table 3
Logistic regression results of Experiment 2 in which Spanish listeners learned to categorize stimuli with relevant variation in one dimension and irrelevant variation in the other dimension without supervision

	Condition 1, duration relevant ($N = 6$)				Condition 2, F1 relevant ($N = 8$)			
	$\mu(\beta)$	$\sigma(\beta)$	Uni	Multi	$\mu(\beta)$	$\sigma(\beta)$	Uni	Multi
Pretest								
Relevant	1.02	1.47	2	0	1.27	1.82	2	2
Irrelevant	1.19	1.35	3		1.79	1.23	4	
Learning phase 1								
Relevant	0.86	1.15	2	1	0.74	1.41	6	0
Irrelevant	0.66	0.37	3		0.18	0.21	1	
Learning phase 2								
Relevant	0.93	1.35	1	1	0.53	0.77	7	0
Irrelevant	0.67	0.36	4		0.13	0.12	0	
Maintenance phase								
Relevant	1.38	1.96	2	0	3.20	2.30	6	0
Irrelevant	0.96	1.25	2		0.68	0.87	2	

The table displays the results of the pretest, learning phases and maintenance phase of Condition 1 (duration relevant) and Condition 2 (formant frequency relevant). The mean β-weights and their standard deviations as well as the number of Listeners using one ("Uni") or both ("Multi") dimensions significantly are shown. Listeners using no dimension significantly are not shown.

Experiments 1 and 2 show that the quantitative differences between supervised and unsupervised learning are considerable. A joint ANOVA with Supervision, Dimension, and Condition as between-subjects variable and Part of the experiment as within-subjects variable should show a significant effect of Supervision. Due to considerable variability in performance between both experiments, and relatively small sample sizes, the difference between supervised and unsupervised learning was not significant ($F\,[1,33] = 0.27$, n.s.) nor were there any relevant interactions.

The results of Experiments 1 and 2 indicate a preference of our Spanish listeners for the dimension of formant frequency. The percentage correct levels in particular showed that performance was better when formant frequency was the relevant dimension. We hypothesize that this is because of the phonological structure of the language, where duration is not important for distinguishing vowels but formant frequency is (Hammond, 2001).

We next tested listeners whose phonology differed from that of the Spanish listeners in Experiments 1 and 2; in Experiment 3, Condition 2 of Experiment 1 was repeated with speakers of American English. While duration may not strictly speaking be a phonemic cue in American English, as described above there is significant variation across vowels in average duration (Hillenbrand et al., 1995), and listeners use duration for distinguishing voicing (Lisker, 1978) as well as in discriminating between tense and lax vowels (Nearey, 1989). No such differentiation is necessary in Spanish, which has only tense vowels (Hammond, 2001). Better performance by the American listeners than the Spanish listeners would indicate effects of the native phonological system in learning new phonetic categories.

4. Experiment 3

To investigate the influence of the native phonology, Experiment 3 compared supervised learning performance of Spanish and American English listeners of a category distinction based on duration.

4.1. Method

4.1.1. Subjects

Ten undergraduate students from the University of Wisconsin, Madison, all native speakers of American English, participated in the experiment and were paid for their participation. None spoke another language besides English, and all reported normal hearing. The post-experiment questionnaire again revealed that all listeners judged all the sounds to be vowels.

4.1.2. Stimuli and procedure

The stimuli were identical to those in Condition 1 of Experiments 1 and 2; duration was the relevant dimension for categorization, and formant frequency varied irrelevantly. The procedure was similar to that of Experiment 1: a pretest, two learning phases and a maintenance phase.

After the listeners had received instructions and signed consent forms, they were seated in a soundproof booth. The pretest and the maintenance phase were identical: subjects were asked to categorize the stimuli into two groups. In the pretest this was done spontaneously, while in the maintenance phase subjects had to try to maintain the response pattern they had discovered in the learning phase. In the learning phase listeners assigned sounds to one of two buttons. If a sound was assigned correctly, a light above the button would light up. If a sound was not assigned correctly, the light belonging to the other button would light up, giving the listener trial-by-trial feedback about the correct response. Listeners were asked to categorize correctly as many stimuli as they could with the feedback given. In the learning phase, 112 stimuli from each category were again presented twice, resulting in 448 trials. After the experiment, participants filled out a questionnaire equivalent to that used in the previous experiments.

4.2. Results and discussion

4.2.1. Signal detection analysis

Again, percent correct and the d' were calculated for each condition and part of the learning phase. Fig. 5 displays these results. All d's differed significantly from zero (minimum $t\,[9] = 6.91$, $p < 0.05$) and all percentages correct were significantly above chance (minimum $t\,[9] = 8.11$, $p < 0.05$), this time with 50% as the expected value since the categories are predefined. We compared the results from Condition 2 from Experiment 1 with those of Experiment 3 in an ANOVA with Language (Spanish versus English) as between-subjects variable and Learning phase (learning phase 1 versus learning phase 2) as within-subjects variable with percent correct and the d' as dependent measures.

There was a significant difference in performance between the two language groups. The performance of English listeners exceeds that of Spanish listeners both in percent correct ($F\,[1,18] = 5.17$, $p < 0.05$) and d' ($F\,[1,18] = 5.45$, $p < 0.05$) in the absence of any significant interactions between Learning phase and Language. The significant main effect of Learning phase ($F\,[1,18] = 8.71$, $p < 0.05$ for both percent correct and $F\,[1,18] = 33.57$, $p < 0.05$ for d') shows that both language groups were able to learn to use the relevant dimension, duration.

Thus categorization based on duration was more successfully achieved by English listeners, who are more familiar with distinguishing vowels based on duration due to their native phonology (Mermelstein, 1978; Whalen, 1989), than by Spanish listeners, who are not as familiar with duration as a cue to vowel category identity (Navarro, 1968).

4.2.2. Logistic regression

As in Experiments 1 and 2, a logistic regression analysis was performed. Table 4 displays the mean β-weights and

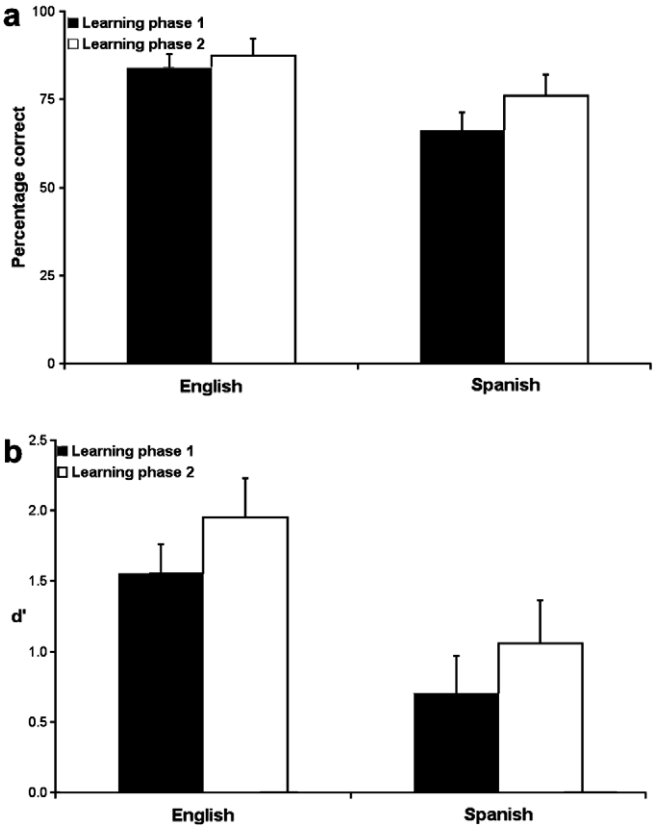

Fig. 5. Percentage correct (a) and d' (b) values for the two learning phases of Experiment 3 (American English listeners), with, for comparison, the equivalent results for the Spanish listeners in Condition 1 of Experiment 1. In both cases duration was the relevant dimension of variation.

standard deviations, as well as the number of subjects using a dimension significantly for each part of the experiment.

Fig. 5 as well as the comparison between Tables 4 and 2 clearly show the differences between the two languages. The mean β-weights for the relevant dimensions were higher for the American English listeners whereas the mean β-weights for the irrelevant dimension formant frequency were higher for the Spanish listeners. These results indicate that using the relevant dimension as well as suppressing an irrelevant one is more feasible when those dimensions play a role in the phonological structure of one's language. A significant interaction between relevance of the dimension and language in an ANOVA with Dimension, Language and Part of the experiment as variables ($F\,[1,18] = 4.55$, $p < 0.05$)[1] warranted separate analyses for the relevant and the irrelevant dimension. For the relevant dimension (duration), there was no significant effect of Language, but for the irrelevant dimension (formant frequency) the β-weights of the Spanish listeners were significantly higher ($F\,[1,18] = 14.49$, $p < 0.05$). This indicates that the Spanish

[1] In fact, all main effects and all interactions were significant except the three-way interaction between Part of the experiment, Dimension, and Language.

Table 4
Results of the logistic regression analysis of Experiment 3 in which American English listeners were trained with supervision to categorize stimuli with relevant variation in one dimension (duration) and irrelevant variation in the other (frequency of the first formant)

	$\mu(\beta)$	$\sigma(\beta)$	Uni	Multi
Pretest				
Relevant	0.30	0.43	3	0
Irrelevant	0.12	0.11	5	
Learning phase 1				
Relevant	1.79	0.05	10	0
Irrelevant	0.95	0.03	0	
Learning phase 2				
Relevant	2.94	0.09	9	0
Irrelevant	1.70	0.06	0	
Maintenance phase				
Relevant	1.16	0.52	9	0
Irrelevant	0.08	0.06	0	

The table shows the β-weights for both duration and frequency of the first formant, their standard deviations as well as the number of listeners significantly using one (“Uni”) or both (“Multi”) dimensions in their categorizations.

listeners experienced difficulty in suppressing the use of formant frequency when it was irrelevant.

A significant effect of Part of the experiment was found for the relevant dimension duration ($F[3,54] = 13.0$, $p < 0.05$) as well as for the irrelevant dimension frequency ($F[3,54] = 3.65$, $p < 0.05$). This effect of Part of the Experiment was modulated by Language in significant interactions for duration ($F[3,54] = 10.4$, $p < 0.05$) and for frequency ($F[3,54] = 3.10$, $p < 0.05$). These interactions point to the known preference of Spanish listeners for formant frequency. We did not find a significant Language effect for the relevant dimension, which is probably due to the high β-weights of the Spanish listeners in the pretest (and, conversely, the low β-weights of the American English listeners in the pretest). When only the learning phases are analyzed with an ANOVA with Language as between-subjects factor and Learning phase (Learning phase 1 and Learning phase 2) as within-subjects variable, there is a significant effect of language for both the relevant dimension ($F[1,18] = 5.46$, $p < 0.05$), where American English has the higher β-weights, and for the irrelevant dimension ($F[1,18] = 7.83$, $p < 0.05$), where Spanish has the higher β-weights.

Taken together, the results of Experiment 3 and Condition 1 of Experiment 1 show the importance of native phonology in learning a new phonetic distinction (see also McAllister et al., 2002 for a similar experiment with non-native listeners with over ten years of exposure to their non-native language). Both Spanish and American English listeners were able to learn a distinction based on duration, but Spanish listeners experienced more difficulty ignoring the irrelevant dimension formant frequency. American English listeners who were more familiar with durational variation in vowels were better able to use this dimension and were also better able to ignore formant frequency.

In Experiments 1–3, learning was limited to situations where one dimension of variation was relevant and another dimension displayed irrelevant variation. This contrasts with the situation in the phonetic inventory of most languages, where it is extremely rare to find truly unidimensional distinctions; there is usually more than one relevant dimension of variation (Lisker, 1978). Furthermore, provided they are detectable, almost all aspects of the speech signal are considered relevant for phonetic categorization (Diehl and Kluender, 1987). So, attending to multiple relevant dimensions is something experienced listeners do continuously and it would be extremely important to be able to do this when acquiring new phonetic categories (Flege and Hillenbrand, 1986). In Experiment 4, we investigate supervised learning of a multidimensional category distinction, exploiting the same dimensions of variation as in the previous experiments, duration and formant frequency. For listeners to obtain a high percentage correct in their categorizations here, both dimensions had to be used in distinguishing the categories.

5. Experiment 4

Experiment 4 investigated supervised learning of a multidimensional category distinction (i.e. both duration and frequency were relevant) in American English listeners.

5.1. Method

5.1.1. Subjects

Eighteen undergraduate students drawn from the same University of Wisconsin subject pool participated in the experiment. All were native speakers of American English (and thus should be able to use both duration and formant frequency in their categorizations). They were paid for their participation. None of the subjects spoke another language besides English and all reported normal hearing. The results of the post-experiment questionnaire were as in the previous experiments: all listeners judged the stimuli to be vowels or extremely like vowels. We chose native speakers of American English because they are familiar with both dimensions. (Spanish speakers, as our experiments showed, have difficulty with duration.) Unfamiliarity with one of the dimensions should thus not be a factor in learning performance.

5.1.2. Stimuli and procedure

Stimulus construction was as in Experiment 1, except that the categories now had two relevant dimensions of variation (duration and formant frequency). Table 5 lists the stimulus characteristics of the learning phase. The pretest and maintenance stimuli were identical to those of Experiment 1, 2, and 3.

The means of the two categories corresponded approximately to the Dutch vowels /y/ and /ø/ as in the Dutch words “fuut” (/fyt/) and “feut” (/føt/). Both frequency of the first formant (formant frequency) and the duration of

Table 5
Stimulus properties of the multidimensional learning stimuli of Experiment 4

Category A "/ø/" as in *feut*			Category A "/y/" as in *fuut*		
Mean	σ	ρ	Mean	σ	ρ
51.8 D	1.22 D		50.4 D	1.21 D	
158 ms	45.1 ms		113 ms	33.4 ms	
		−0.95			−0.95
9.9 ERB	1.32 ERB		8.16 ERB	1.33 ERB	
441.6 Hz	96.1 Hz		327.6 Hz	78.7 Hz	

The duration in DUR (and ms) and formant frequency in ERB and their respective standard deviations are presented for both categories. The pretest and maintenance stimuli are identical to those used in Experiment 1 and can be found in Table 3.

the sound (duration) were varied in creating the categories: /y/ is shorter and has a lower F1 than /ø/ (see the fourth panel of Fig. 2).

The procedure was as in Experiment 3: a pretest, two learning phases and a maintenance phase. In the pretest and maintenance phases subjects categorized the stimuli into two groups, in the pretest choosing labels as they wished, but in the maintenance phase trying to maintain the rule they had discovered in the learning phases. In the learning phases trial-by-trial feedback was again provided by lights above the response buttons.

5.2. Results and discussion

5.2.1. Signal detection analysis

Fig. 6 shows mean percentage correct and mean d' for the first and second learning phase of Experiment 4. The percentage correct and the d' differed significantly from their respective chance levels (50% and 0) in all phases (minimum $t\,[17] = 6.10$, $p < 0.05$), but the difference between the first and second phase in the figures does not give a strong indication for a learning effect. Two ANOVAs with Learning phase as within-subject variable and percentage correct or d' as dependent variables showed no significant effect for either percentage correct ($F\,[1,17] = 0.90$, n.s.) or d' ($F\,[1,17] = 0.30$, n.s.).

While the signal detection measures presented no evidence of learning over time, both differed significantly from chance levels, indicating that listeners were sensitive to the distributional information and the trial-by-trial feedback they received.

5.2.2. Logistic regression

The four panels of Fig. 7 present the β-weights for duration and formant frequency for each listener in each part of the experiment. The abscissa shows the β-weight for duration and the ordinate shows the β-weight for frequency (see Nearey, 1997). Listeners who used both dimensions are identified by asterisks, listeners who used only formant frequency as plus-signs, listeners who used only duration as crosses, and listeners who used neither dimension significantly as circles. Optimal performance corresponds to a point in the upper right hand corner of the Figure, with a φ of 45° (both dimensions are given equal weight) and far from the origin (reflecting consistent behavior).

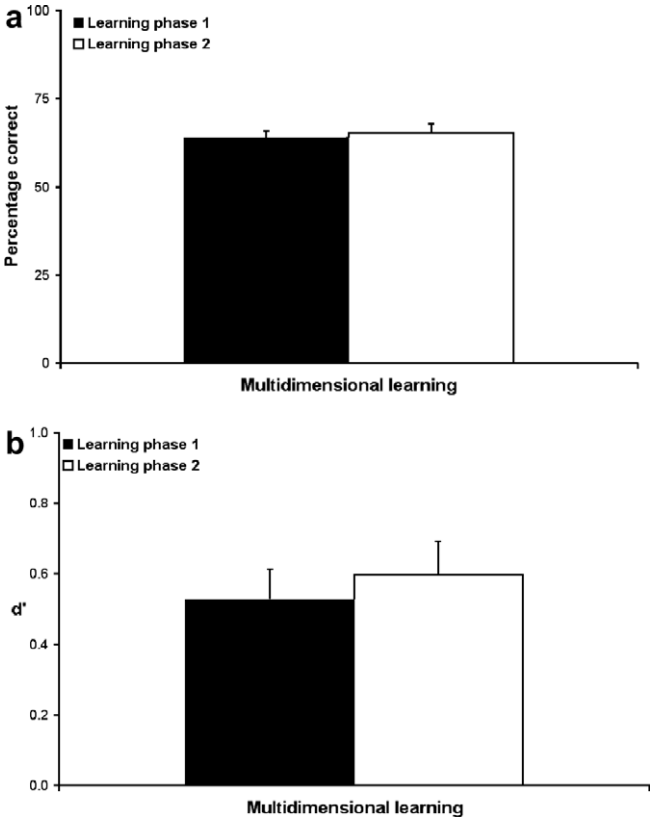

Fig. 6. Percentage correct (a) and d' (b) values for the two learning phases of Experiment 4 (two relevant dimensions of variation).

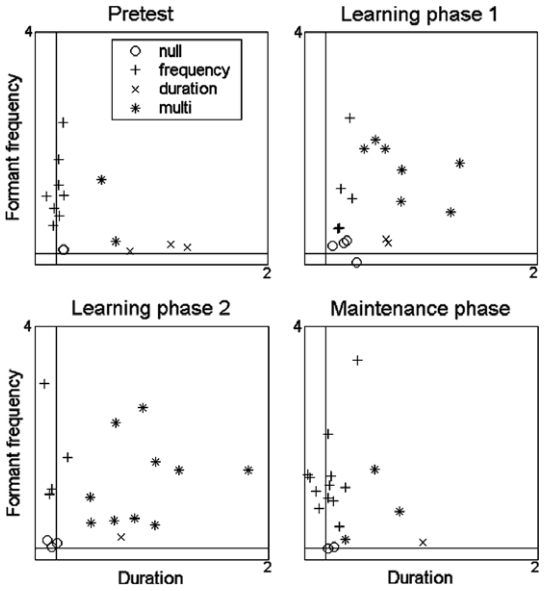

Fig. 7. Scatter plots of individual β-weights for the two dimensions (duration, formant frequency) in Experiment 4 (two relevant dimensions of variation), for each of the four parts of the experiment.

The upper left panel of Fig. 7 shows performance in the pretest. The majority of the listeners had a preference for a unidimensional solution with frequency (plus signs), which had also been the case in the pretest of Experiment 3. The upper right and lower left panel show the learning phases. Over time, the number of listeners using both dimensions in categorization increases (more asterisks), as does their consistency (asterisks further from the origin). In the maintenance phase, when feedback was no longer given, much of this learning seems to be lost and the number of listeners using only formant frequency as the relevant dimension is even larger than in the pretest.

Most subjects succeeded in reliably using one or more dimensions, although some failed to use any dimension significantly. It would be desirable to have a measure of the majority's central tendency and variability, because 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 subjects used duration exclusively, and the others formant frequency, the average β-weights might both exceed chance suggesting that participants on average used both dimensions, even though no individuals did so. A measure that integrates performance on both dimensions would therefore be useful.

Here, we derive such a measure by computing the angle formed by the line connecting each subject's β-weight to the origin, on a graph where the x-axis represents duration,and the y axis frequency (as in Fig. 7), and also computing the length of this line. These computations were done by transforming the Cartesian coordinates of the β-weights for duration and 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:

$$A = \sqrt{\beta_{\text{dur}}^2 + \beta_{\text{freq}}^2} \tag{1}$$

$$\theta = (\beta_{\text{freq}}/\beta_{\text{dur}}) \quad \text{if } \beta_{\text{dur}} < 0 \tag{2a}$$

$$\theta = (\beta_{\text{freq}}/\beta_{\text{dur}}) + \pi \quad \text{if } \beta_{\text{dur}} > 0; \quad \varphi - 2\pi \quad \text{if } \varphi > \pi \tag{2b}$$

In our analysis, φ ranges between π and $-\pi$ radians. When φ equals $1/2\,\pi$, listeners purely use frequency, when φ equals 0, listeners use only duration, but when φ is close to $1/4\,\pi$ performance lies between those two angles, i.e. both duration and frequency are used. As can be seen from Fig. 7, most listeners fall in the upper right plane, between 0 and $1/2\,\pi$.

The other polar coordinate, A, ranges between zero and plus infinity. A large A indicates that a subject was internally consistent (though a large average A over subjects need not reflect consistent weights of each dimension), while a small A indicates that listeners' categorizations tend not to be internally consistent. In Fig. 7, the listeners who categorized using both dimensions (the asterisks) are farther removed from the origin, while listeners who use no dimension significantly (the circles) are all very close to the origin. Table 6 displays the mean values for φ, A and their standard deviations as well as the number of listeners (total $N = 18$) using one or two dimensions significantly.

The central question is whether the mean φ of each learning phase differed significantly from 0 (representing a unidimensional duration solution) and from $1/2\,\pi$ (representing a unidimensional formant frequency solution). This was tested with two t-tests corrected for the increased type I error with Bonferroni correction for every phase of the experiment. This resulted in significant differences with both 0 and $1/2\,\pi$ in all phases (min $t\,[17] = 2.47$, all $p < 0.05$). The average φ was multidimensional even though not all subjects categorized using a multidimensional rule (subjects using formant frequency canceled out those using duration). Importantly, the number of listeners preferring the multidimensional solution over a unidimensional one increased during the learning phases, showing the ability of our listeners to profit from trial-by-trial feedback and distributional information. Flege et al. (1997) show the ability of Spanish listeners to use either duration or frequency (although seldom both simultaneously) in identifying vowels.

The consistency measure A was statistically evaluated in an ANOVA with Part of the experiment as within-subject variable. As with the signal detection measures, the different phases of the experiment did not differ significantly from each other ($F\,[3, 51] = 0.784$, n.s.).

A final interesting comparison is that between unidimensional supervised learning and multidimensional supervised learning by our American English listeners. Percentage correct and d' were analyzed in a joint ANOVA with Part of the experiment as within-subjects factor and Experiment (unidimensional versus multidimensional) as between-subjects factor. Performance in the unidimensional learning experiment was consistently better for both percentage

Table 6
Results of the logistic regression analysis of Experiment 4 in which English listeners were trained with supervision on a category distinction where both dimensions were relevant

	φ (σ)	A (σ)	Uni	Multi
Pretest				
Duration	0.27 (0.28)	1.08 (0.67)	3	4
F1			9	
Learning phase 1				
Duration	0.22 (0.20)	0.96 (0.77)	2	7
F1			3	
Learning phase 2				
Duration	0.34 (0.28)	1.26 (0.89)	1	10
F1			4	
Maintenance phase				
Duration	0.35 (0.24)	1.07 (0.77)	1	3
F1			12	

The angle φ, the consistency measure A as well as their respective standard deviations are shown, as well as the number of listeners significantly using one ("Uni") or both ("Multi") dimensions in their categorizations. Listeners using no dimension are not shown ($N = 18$).

correct ($F[1,26] = 24.67$, $p < 0.05$) and d' ($F[1,26] = 31.14$, $p < 0.05$).

Experiment 4 thus showed that listeners were sensitive to the distributional information and trial-by-trial feedback provided to them in multidimensional category learning. However, performance in Experiment 4 was considerably worse than in Experiment 3. Learning a category distinction with more than one relevant dimension was significantly more difficult than learning to use one dimension while simultaneously learning to ignore the other.

The amount of exposure our listeners received (448 stimuli) was substantial for such experiments, but is insignificant compared to the exposure received by infants, or by adults learning a second language. Despite this relatively small amount of exposure, more than half of the listeners were able to use both dimensions after the learning phase. The loss of this ability in the maintenance phase that followed within minutes is striking. Listeners almost invariably prefer unidimensional solutions in category learning (Ashby et al., 1999). But although the learning phases of Experiment 4 showed that this preference can be modified, without distributional information and trial-by-trial feedback listeners reverted to unidimensional categorization. An explanation for this phenomenon is that our listeners may be extremely sensitive to the removal of distributional information favoring one dimension above the other. In the absence of this information in the maintenance phase, listeners adjusted their categorization tendencies to suit. Goudbeek et al. (in preparation), in a study with Dutch listeners, used a maintenance phase where feedback was absent, but distributional information remained present; their listeners were able to maintain the learned multidimensional categorization strategy in such a maintenance phase, consistent with the hypothesis that successful categorization depends upon the availability of the distributional information.

6. General discussion

The acquisition of speech categories by adult listeners is sensitive to the distributional structure of the input, is affected by the categories of the native phonemic inventory, and is greatly facilitated by the provision of intensive feedback.

The stimuli in our four experiments displayed tightly controlled variation in dimensions known to be important in speech perception: duration and formant frequency. Depending on condition, each type of variation was either relevant or irrelevant to the category distinction listeners were expected to learn. The first three experiments examined the acquisition of a unidimensional categorization distinction. In Experiment 1, Spanish listeners categorized non-native speech sounds with the aid of trial-by-trial feedback (supervision). The results showed that listeners could indeed learn to attend to a relevant dimension while suppressing an irrelevant one. For these listeners, however, learning and maintaining a distinction based on formant frequency appeared easier than learning and maintaining a distinction based on duration. In Experiment 2, listeners from the same population categorized the same stimuli, but without trial-by-trial feedback, i.e., on the basis of distributional information alone. Performance was worse than in Experiment 1, and was even more strongly affected by which dimension was relevant. Nonetheless, on several measures performance did differ from chance, showing that listeners were sensitive to the distributional structure of the input. In Experiment 3, American English listeners were presented with the duration-relevant stimuli of Experiments 1 and 2, with trial-by-trial feedback during learning. These listeners, who are more familiar than Spanish listeners with durational variation in their native vowels, were better able than the Spanish listeners to acquire the duration-based contrast.

Finally, in Experiment 4 listeners learned a category distinction with two relevant dimensions. Although the American English listeners presented with this task were acquainted with both dimensions involved, and although they received trial-by-trial feedback on their categorizations, performance was considerably impaired compared to equivalent learning of a unidimensional distinction. Nevertheless, listeners were certainly sensitive to the distributional information provided and the majority of them learned to make use of both dimensions in the categorization.

Across all the experiments there has been a consistent appearance of sensitivity to the distributional structure of the input. All listeners reported (in the post-experiment questionnaires) that the stimuli they had heard sounded like speech sounds. Yet the sounds were not like the vowels of either of the native languages involved, and the distributional structure of the category distinction represented in the input was unfamiliar to these listeners. Despite this, they acquired the category distinction at least to some degree; this is quite obvious from the clear improvement across the learning phases visible in Figs. 3 and 5, and even evident in Figs. 4 and 6 for the harder cases, the unsupervised learning of Experiment 2 and the multidimensional categorization of Experiment 4. Tables 2–4 and 6 likewise show that improvement in performance across the experiment was consistent across listeners.

The better performance in Experiments 1 and 3 versus Experiment 4 additionally shows that learning to identify a category structure with one relevant dimension of variation and one irrelevant dimension of variation is more feasible than learning to identify a category structure with two relevant dimensions of variation. Nevertheless, even with as little as a few hundred presentations, listeners proved sensitive to the distributional information available to them in Experiment 4. Learning to integrate two dimensions to distinguish two phonetic categories is difficult (in line with previous findings of Flege and Hillenbrand, 1986), but not impossible. In summary, our results have shown that it is certainly possible for adult listeners already

in possession of a phonological system to attend systematically to the distributional structure of auditory input and learn a new vowel categorization from it.

Notwithstanding these achievements, clear effects of the native phonology were also manifest in our findings. Spanish listeners found it much easier to acquire a category distinction based on frequency than one based on durational variation; Figs. 3 and 4 make this main effect across conditions abundantly clear. Tables 2 and 3 likewise show that a preference for the use of frequency was evident in the pretest and maintenance phases when no distributional cues were present. Frequency distinctions between speech segments are familiar to these listeners; durational distinctions are not. For American listeners, durational distinctions between vowels are more familiar; listeners use duration to distinguish voicing contrasts (Lisker, 1978; Flege and Hillenbrand, 1986), and there are further relevant effects such as the distinction between tense and lax vowels in English which involves (allophonic) duration differences between longer tense vowels and shorter lax vowels (Nearey, 1989; Smiljanić and Bradlow, 2005). As Kawahara (2006) has shown with Japanese and English listeners, the duration of auditory stimuli can be perceived differently by listeners with differing phonologies. Given their experience, the American listeners' performance with exactly the same durationally based categorization was far better than that of the Spanish listeners, as the comparison in Fig. 5 again makes clear. Thus the ability to acquire a novel vowel categorization from auditory input is modulated by the degree to which the categorization maps to the type of distinctions required by the native phonology. A distinction requiring the use of dimensions deemed irrelevant by the native phonology is harder to acquire than a distinction based on the application of sources of information which are exploited by the native phonology too, so that using them is a familiar task.

Note that we did not find inhibitory effects of the native vowel categories themselves in our studies. For both listener populations we tested, the Dutch vowels /y/, /ʏ/ and /ø/ on which the stimuli were based are unfamiliar and fall in an empty portion of the native vowel space. However, the vowel space of American English is far more densely populated than the vowel space of Spanish; in consequence, interference of native categories might have affected our American more than our Spanish listeners. If that was so, however, then the interference was clearly not of a kind that inhibited their ability to acquire the categories they were exposed to.

Effects of native categories on the acquisition of non-native categories may also be facilitatory. As noted in the Introduction, there are three possibilities which PAM (Best, 1995) allows for when non-native speech categories are categorized within a native phonological system. First, the new categories might both map to one and the same native phoneme; then, discrimination is very difficult, a well-known example being English /r/ and /l/ which map to a single native Japanese phonetic category. But second, the two non-native phonemes might map (well or imperfectly) to separate native phonetic categories. The native and non-native categories do not have to be identical; as long as there is a sufficiently consistent mapping between the two sets of categories, discriminating the non-native sounds is easy. Third, one of the two categories may map better than the other to a native phonetic category; in this case, non-native category learning depends on the relative goodness of fit of each non-native category to the native category, with a large difference in fit making discrimination and non-native category learning easier. Thus two of these three possibilities allow mapping to the native category system to assist the acquisition of a non-native distinction. However, we have no evidence for such facilitatory effects in our data either. Judging by the IPA vowel charts (Gussenhoven, 1999; Ladefoged, 1999), the closest American English vowel to any of the three Dutch vowels is the vowel /ʊ/, which is closer to /y/ than to /ʏ/ or /ø/. If this difference in proximity had exercised effects, it should have been to make acquisition of the distinction between /y/ and /ʏ/ in Experiment 4 somewhat easier, but there was little sign of this categorization being at all easy for our American listeners. The effects of native phonology which we observed were confined to the effects of the type of information – duration versus formant frequency – on which the categorization distinctions are based.

A further effect in our data concerns the role of feedback, or supervision. Comparison of Experiment 1 versus Experiment 2 suggests that supervised learning produces higher levels of achievement than unsupervised learning, even when the distinction to be learned involves only one dimension of variation. The acquisition of unfamiliar auditory category distinctions is clearly not an easy task, and feedback on performance trial-by-trial is useful to learners. Note that there was no differential effect of the availability of feedback for the easier versus the harder distinction in this task, suggesting that listeners do not restrict themselves to exploiting such assistance only when faced with a distinction based on a type of information irrelevant to native phonological contrasts.

The present results have significantly extended our knowledge of the factors which control the success of speech category acquisition in its own right, separately from the role such acquisition must necessarily play in the learning of a complete phonological system. To acquire novel auditory categories, listeners simultaneously draw on the structure of the input they receive and on their existing linguistic knowledge. Thus the native phonological system plays a role not only when a whole new repertoire is being acquired; any newly acquired individual distinction can be affected by existing phonological knowledge. The American English participants in our studies had no prior exposure to Dutch, and presumably had no particular motivation to acquire Dutch phonemic distinctions, but nonetheless they performed significantly above chance in acquiring the Dutch vowel distinctions, and, presumably by drawing on their native experience with the use

of vocalic variation in duration, were even able to perform better than these Spanish participants who might be considered as nominally motivated to acquire knowledge of the Dutch language (given that they were exchange students).

We view the role of the native phonology in speech category acquisition as not in principle different from any other body of existing knowledge. It is simply the most relevant existing knowledge that the listener has. Learning processes are well adapted for drawing on existing knowledge to fine-tune perceptual performance; this can be seen in the use of lexical knowledge to fine-tune the native phonemic categorizations (Norris et al., 2003). This fine-tuning is held to be the process underlying listeners' rapid adaptation to new talkers (Eisner and McQueen, 2005), and it is certainly especially well adapted to the complexity of speech perception, in which highly complex signals consist of components which are not clearly demarcated. Nonetheless, the same kind of fine-tuning can be drawn upon in the perception of printed text (Norris et al., 2006), where clearly demarcated components would seem to make it unnecessary. Norris et al. (2006) argue that the ability to use information from one level of analysis to inform learning about perceptual distinctions at another level has such power that it cannot be restricted to learning which would be impossible without it. In the present case, effects of the native phonology were exercised even though the acquisition task was not explicitly one of vowel learning.

The final question is the applicability of our results to learning of natural auditory categories in general, given that the factors affecting success in the learning of non-native speech categories are presumably not restricted to this situation. Supervision is usually not available to learners, certainly not on a case-by-case basis. Our results suggest that the role of existing knowledge – the native phonology, in this case – is larger if no supervision is provided. Without trial-by-trial feedback, listeners in Experiment 2 experienced more difficulty in ignoring the dimension relevant in their native phonology, even though the distributional properties of the stimuli indicated otherwise. They preferred to use the dimension better known to them, in this case formant frequency. We would suggest that the provision of supervision amounts to the provision of a very reliable knowledge source. That is, human learning is extremely efficient and makes use of the best information available. This may be supervisory feedback, or it may be existing phonological knowledge. As the literature on non-native phoneme learning repeatedly demonstrates, the use of existing phonological knowledge sometimes does not help; this simply underlines the fact that among category acquisition tasks, the acquisition of non-native speech categories is one of the most difficult.

Acknowledgements

The research reported in this paper was supported by a Max Planck Society doctoral scholarship. Experiments 3 and 4 were made possible by a Dutch Scientific Council (NWO) travel grant, and Experiment 1 was supported by the NWO-SPINOZA project "Native and Non-Native Listening" to A.C. We thank Keith Kluender, University of Wisconsin, Madison, for financial assistance and consultation, Anita Wagner and Laurence Bruggeman for help in recruiting and testing the Spanish listeners, and two reviewers for helpful suggestions on the manuscript. M.G. is now at The Geneva Emotion Research Group of the Department of Psychology, University of Geneva.

References

Agresti, A., 1990. Categorical Data Analysis. John Wiley and Sons, Inc., New York.

Aoyama, K., Flege, J.E., Guion, S.G., Akahane-Yamada, R., Yamada, T., 2004. Perceived phonetic dissimilarity and L2 speech learning: the case of Japanese /r/ and English /l/ and /r/. J. Phon. 32, 233–250.

Ashby, F.G., Maddox, W.T., 1993. Relationships between prototype, exemplar, and decision bound models of categorization. J. Math. Psychol. 37, 372–400.

Ashby, F.G., Alfonso-Reese, L.A., Turken, A.U., Waldron, E., 1998. A neuropsychological theory of multiple systems in category learning. Psychol. Rev. 105, 442–481.

Ashby, F.G., Queller, S., Berretty, P., 1999. On the dominance of unidimensional rules in unsupervised categorization. Percept. Psychophys. 61, 1178–1199.

Best, C.T., 1994. The emergence of native-language phonological influences in infants: a perceptual assimilation model. In: Goodman, J.C., Nusbaum, H.C. (Eds.), The development of speech perception. The MIT Press, Cambridge, MA, pp. 167–224.

Best, C.T., 1995. A direct realist view of speech cross language speech perception. In: Strange, W. (Ed.), Speech Perception and Linguistic Experience: Issues in Cross-Language Research. New York Press, Baltimore, MD, pp. 171–206.

Best, C.T., Tyler, M.D., 2007. Nonnative and second language speech perception: commonalities and complementaries. In: Munro, M.J., Bohn, O.-S. (Eds.), Second Language Speech Learning: The Role of Language Experience in Speech Perception and Production. John Benjamins, Amsterdam, pp. 13–34.

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. J. Exp. Psychol.: Hum. Percept. Perform. 14, 345–360.

Boersma, P., Weenink, D., 2003. Praat 4.1 [Computer software]. Retrieved from [URL] <www.fon.hum.uva.nl/praat/>.

Booij, G., 1995. The Phonology of Dutch. Clarendon Press, Oxford.

Bradlow, A.R., 1995. A comparative study of English and Spanish vowels. J. Acoust. Soc. Amer. 97, 1916–1924.

Broersma, M., 2006. Learning to ignore a perceptual cue: nonnative listeners outperform native listeners. Poster presented at the 151st Meeting of the Acoustical Society of America, Providence. RI. J. Acoust. Soc. Amer. 119, 3270–3271 (Abstract).

Burnham, D.K., Earnshaw, L.J., Clark, J., 1991. Development of categorical identification of native and non-native bilabial stops: infants, children and adults. J. Child Lang. 18, 231–260.

Cutler, A., Broersma, M., 2005. Phonetic precision in listening. In: Hardcastle, W., Beck, J. (Eds.), A Figure of Speech: A Festschrift for John Laver. Erlbaum, Mahwah, NJ, pp. 63–91.

Diehl, R.L., Kluender, K.R., 1987. On the categorization of speech sounds. In: Harnad, S. (Ed.), Categorical Perception. Cambridge University Press, Cambridge, pp. 226–253.

Eisner, F., McQueen, J.M., 2005. The specificity of perceptual learning in speech processing. Percept. Psychophys. 67, 224–238.

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

Flege, J., 1989. Differences in inventory size affect the location but not the precision of tongue positioning in vowel production. Lang. Speech 32, 123–147.

Flege, J.E., 1992. The intelligibility of English vowels spoken by British and Dutch talkers. In: Kent, R. (Ed.), In: Intelligibility in Speech Disorders: Theory, Measurement, and Management. John Benjamins, Amsterdam, pp. 157–232.

Flege, J.E., 1995. Second language speech learning: theory, findings and problems. In: Strange, W. (Ed.), Speech Perception and Linguistic Experience: Theoretical and Methodological Issues. York Press, Timonium, MD, pp. 233–273.

Flege, J.E., Hillenbrand, J., 1986. Differential use of temporal cues to the /s/-/z/ contrast by native and non-native speakers of English. J. Acoust. Soc. Amer. 79, 508–517.

Flege, J.E., Bohn, O., Jang, S., 1997. Effects of experience on non-native speakers' production and perception of English vowels. J. Phon. 25, 437–470.

Francis, A.L., Nusbaum, H.C., 2002. Selective attention and the acquisition of new phonetic categories. J. Exp. Psychol.: Hum. Percept. Perform 28, 349–366.

Francis, A.L., Baldwin, K., Nusbaum, H.C., 2000. Effect of training on attention to acoustic cues. Percept. Psychophys. 62, 1668–1680.

Goudbeek, M., Swingley, D., Smits, R., in preparation. Supervised and unsupervised learning of acoustic categories.

Gussenhoven, C., 1999. Illustrations of the IPA: Dutch. Handbook of the International Phonetic Association. Cambridge University Press, Cambridge, pp. 74–77.

Gussenhoven, C., in press. Vowel duration, syllable quantity and stress in Dutch. In: Kristin Hanson, Sharon Inkelas (Eds.), The Nature of the Word: Essays in Honor of Paul Kiparsky. MIT Press, Cambridge, MA.

Hammond, R.M., 2001. The sounds of Spanish: Analysis and application. Cascadilla Press, Sommerville, MA.

Hillenbrand, J., Getty, L.A., Clark, M.J., Wheeler, K., 1995. Acoustic characteristics of American English vowels. J. Acoust. Soc. Amer. 97, 3099–3111.

Homa, D., Cultice, J.C., 1984. Role of feedback, category size, and stimulus distortion on the acquisition and utilization of ill-defined categories. J. Exp. Psychol. – Learn. Mem. Cogn. 10, 83–94.

Kawahara, S., 2006. Contextual effects on the perception of duration. Poster presented at the 151st Meeting of the Acoustical Society of America, Providence, RI. J. Acoust. Soc. Amer. 119, 3243 (Abstract).

Ladefoged, P., 1999. American English. Handbook of the International Phonetic Association. Cambridge University Press, Cambridge, pp. 41–44.

Lisker, L., 1978. Rapid versus rabid: a catalogue of acoustic features that may cue the distinction. Haskins Laboratories Status Report in Speech Research SR-54, pp. 127–132.

Logan, J.S., Lively, S.E., Pisoni, D.B., 1991. Training Japanese listeners to identify English /r/ and /l/: a first report. J. Acoust. Soc. Amer. 89, 874–886.

Love, B.C., 2002. Comparing supervised and unsupervised category learning. Psychon. Bull. Rev. 9, 829–835.

Macmillan, N.A., Creelman, C.D., 1991. Detection Theory: A User's Guide. Cambridge University Press, New York.

Maye, J., Gerken, L., 2000. Learning phoneme categories without minimal pairs. In: Proc. 24th Annual Boston University Conf. on Language Development, pp. 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.), Proc. 25th Annual Boston University Conf. on Language Development, pp. 480–490.

McAllister, R., Flege, J.E., Piske, T., 2002. The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English and Estonian. J. Phon. 30, 229–258.

Mermelstein, P., 1978. On the relationship between vowel and consonant identification when cued by the same acoustic information. Percept. Psychophys. 23, 331–336.

Navarro, T., 1968. Studies in Spanish phonology. University of Miami Press, Coral Gables.

Nearey, T.M., 1989. Static, dynamic, and relational properties in vowel perception. J. Acoust. Soc. Amer. 85, 2088–2113.

Nearey, T.M., 1997. Speech perception as pattern recognition. J. Acoust. Soc. Amer. 101, 3241–3254.

Norris, D., McQueen, J.M., Cutler, A., 2003. Perceptual learning in speech. Cogn. Psychol. 47, 204–238.

Norris, D., Butterfield, S., McQueen, J.M., Cutler, A., 2006. Lexically guided retuning of letter perception. Q. J. Exp. Psychol. 59, 1505–1515.

Nosofsky, R.M., 1990. Exemplar-based approach to categorization, identification and recognition. In: Ashby, F.G. (Ed.), Multidimensional Models of Perception and Cognition. Lawrence Erlbaum Associates, New York, pp. 363–393.

Pierrehumbert, J., 2003. Phonetic diversity, statistical learning, and acquisition of phonology. Lang. Speech 46, 115–154.

Shepard, R., Hovland, C., Jenkins, H., 1961. Learning and memorization of classifications. Psychol. Monogr. 75, 1–42.

Smiljanić, R., Bradlow, A.R., 2005. Production and perception of clear speech in Croatian and English. J. Acoust. Soc. Amer. 118, 1677–1688.

Smits, R., Warner, N., McQueen, J., Cutler, A., 2003. Unfolding of phonetic information over time: a database of Dutch diphone perception. J. Acoust. Soc. Amer. 113, 563–574.

Strange, W., 1995. Speech Perception and Linguistic Experience: Issues in Cross-language Speech Research. York Press, Timonium, MD.

Whalen, D., 1989. Vowel and consonant judgments are not independent when cued by the same information. Percept. Psychophys. 46, 284–292.