The acquisition of auditory categories

ISBN-10: 90-76203-27-X

ISBN-13: 978-90-76293-27-0

Cover design: Linda van den Akker, Inge Doehring

Cover Illustration: John & Martijn Goudbeek, Thanks to Sara Lee Benelux.

Printed and bound by Ponsen en Looijen bv, Wageningen

The acquisition of auditory categories

Een wetenschappelijke proeve

op het gebied van de Sociale Wetenschappen

PROEFSCHRIFT

ter verkrijging van de graad van doctor

aan de Radboud Universiteit Nijmegen

op gezag van de Rector Magnificus Prof. dr. C.W.P.M. Blom,

volgens besluit van het College van Decanen

in het openbaar te verdedigen

op donderdag 8 maart 2007

om 15:30 uur precies

door

Martijn Bastiaan Goudbeek

geboren op 31 december 1975 te Enschede

PROMOTOR:	Prof. dr. A. Cutler
COPROMOTORES:	Dr. R.L.H.M. Smits
	Dr. D. Swingley
MANUSCRIPTCOMMISSIE:	Prof. dr. L. Boves
	Prof. dr. A.W. Bronkhorst
	Dr. W.T. Maddox

The research reported in this thesis was supported by a grant from the Max-Planck Gesellschaft zur Förderung der Wissenschaften, München, Germany and by an NWO travel grant.

"Forty-two!" yelled Loonquawl. "Is that all you've got to show for seven and a half million years' work?"
"I checked it very thoroughly," said the computer, "and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you've never actually known what the question is."
- Douglas Adams, The Hitchhiker's Guide to the Galaxy..

Ὧν ἡ σοφία παρασκευάζεται εἰς τὴν τοῦ ὅλου βίου μακαριότητα πολὺ μέγιστόν ἐστιν ἡ τῆς φιλίας κτῆσις.
- Epicurus, Principal Doctrines: 27

Voorwoord

Toen ik in 2001 begon aan het onderzoek dat zou leiden tot dit proefschrift, kon ik niet vermoeden dat ik zoveel zou leren en zoveel hulp zou krijgen. Dit proefschrift had niet bestaan zonder Anne Cutler, Roel Smits en Daniel Swingley. Hen wil ik op de eerste plaats hier bedanken.

Roel, met de rust en het vertrouwen dat je uitstraalde heb je er voor gezorgd dat ik nooit twijfelde aan het afronden van het hele zaakje. Je zorgvuldigheid gekoppeld aan je vaardigheid in het nemen van rigoureuze beslissingen, zijn deel geworden van mijn wetenschappelijke houding. Je enthousiasme voor ons onderzoek heeft me vaak over de dode punten heen geholpen.

Dan, the emphasis you put on how best to tell the story of our studies is still a very valuable lesson to me. You convinced me time and time again that results that seemed extremely boring and useless were interesting and worthwile. This, and your lessons in clear and convincing writing have made me a better researcher.

De combinatie van jullie persoonlijkheden, interesses en kennis heeft uitzonderlijk goed gewerkt voor mij. Steeds vaker vraag ik me af hoe jullie dat voor elkaar hebben gekregen. Dank daarvoor.

Anne, you have been an inspiring mentor and an example of scientific passion. You were as unforgiving of my sloppiness as you were enthusiastic about the final papers. Working with you has become more and more of a pleasure for me (perhaps I get less sloppy). Plus, the comprehension group would not be such a nice working environment were it headed by someone else. Somehow you have succeeded in

combining the indiviualism inherent in doing research with a very pleasant and cohesive group structure. Thank you.

Daarnaast hebben een paar anderen een belangrijke bijdrage geleverd aan dit proefschrift. Zonder mijn paranimfen Anita Wagner en Keren Shatzman had ik de laatste stappen naar de promotie niet zo makkelijk gezet. Mirjam Broersma, onlosmakelijk verbonden met mijn tijd op het MPI, heeft de Nederlandse samenvatting gevrijwaard van taal- en spelfouten.

Tot slot, mijn vrienden, vriendinnen en familie. Behoudens het af en toe indrukken van knopjes als proefpersoon hebben jullie geen inhoudelijke bijdrage aan het proefschrift geleverd. Waarvoor hulde. Dankzij jullie bleef 'de wereld' iets echts en leuks en tastbaars. Jullie ben ik veel en veel meer verschuldigd dan dank voor het schrijven van dit proefschrift.

Martijn Goudbeek

December 2006

Contents

Chapter 1

Introduction

Categorization is a fundamental cognitive process. It enables us to structure an otherwise unstructured world. With categorization, the continuous stream of perceptions becomes a series of separate and discrete ones with each percept labeled as belonging to a category. Categorization and category learning are involved in a surprisingly large number of cognitive processes. Without categories, we would be unable to recognize the colors of the rainbow, to appreciate tonal music, to talk about kinds of animals at the zoo, to recognize friends or to read someone's handwriting. In other words, categorization is present in all situations where previous experience guides our present interpretations. Understanding spoken language is a prime example of such a situation. It involves categorization at multiple levels, ranging from recognizing consonants and vowels to interpreting grammatical structures and context. In this thesis, we consider the process of acquiring the ability to categorize speech sounds, in other words, the learning of phonetic categories.

The first learning of phonetic categories takes place in infancy, as we tune in to the linguistic sounds around us. Later in life, if we try to acquire a second language, we again need to master new sounds. This time the sounds may be sounds that do not occur in our native language. Although a lot is known about the abilities of infants to discriminate native and non-native phonemes, exactly how they develop these abilities is not well understood. An important observation is that because

listening abilities *precede* speaking abilities, infant learning of phonetic categories cannot be explicitly supervised. Supervision, defined as feedback on performance, can only be given when there is something to be (positively or negatively) reinforced. Without observable language behavior in infants, this feedback is difficult to imagine. From there, the idea follows that infant phonetic category learning must involve some sort of statistical pattern recognition (Saffran, Aslin, & Newport, 1996; Jusczyk 1997; Lotto, 2000). This distinction between *supervised* and *unsupervised* learning of phonetic categories is central to this thesis.

An important difference between first language learning by infants and second language learning by adults, therefore, is the availability of feedback. In adult learning, involving mature speaking and listening abilities, there is at least the possibility of supervision in the form of feedback on verbal or non-verbal responses. This thesis examines both kinds of learning. In all experiments, the presence or absence of feedback will be manipulated alongside the kind of probabilistic information contained in the category structures presented to the listeners.

Categorization and category learning

Three closely related but fundamentally different cognitive processes need to be distinguished: *categorization*, *identification*, and *discrimination*. Categorization is the mapping of many stimuli varying along continuous dimensions to a (usually much smaller) set of categories where all the members that fall into the same category are interpreted as being equivalent (Nosofsky, 1986, 1990). For example, humans map a wide range of wavelengths of the electromagnetic spectrum to the same color name,

and the number of colors is limited compared to the range of wavelengths the human eye can perceive. (400 to 700 nm).

In contrast to categorization, identification is the one to one mapping of stimulus and label. Face recognition, for instance, involves identifying family members by their facial features. Recognizing the gender of individuals by their facial features, on the other hand, involves categorization.

Discrimination involves processing whether stimuli are the same or different and lies at the basis of both categorization and identification. If we were unable to reliably discriminate two stimuli we could not reliably assign them to two different categories either. All colors would merge to one and all family members would effectively have the same face. This is the sad lot of brain damaged patients who are suffering from proposagnosia: the inability to recognize faces (Hecean & Anelergues, 1962). Our ability to discriminate is limited, however. The difference between electromagnetic radiation with a wavelength of 450 nm and electromagnetic radiation with a wavelength of 451 nm, for example, is impossible to discriminate for a human.

In categorization, there are two kinds of categories: conceptual categories and perceptual categories (Medin & Barsalou, 1987). *Conceptual* categories are usually defined in language terms and are differentiated from each other by discrete features: "round head versus oval head" or "long versus short legs" (see for example Minda & Smith, 2001, 2002). Examples include concepts like mammal, bird, politics, and democracy. These conceptual categories are not the subject of this thesis, although sometimes the theory concerning them does come into play. *Perceptual* categories are defined in psychophysical terms (they map the physical world onto the psychological) and have continuous dimensions rather than discrete features.

Examples include color categories, phonetic categories, and faces. Although both category conceptions are sometimes assumed to be two ends of a continuum (Goldstone & Barsalou, 1998) and to share a common background in terms of similarity (Gureckis & Love, 2003), the processes involved in learning conceptual categories may differ from those involved in learning perceptual categories. Irrespective of the validity of the difference, here, the learning of perceptual and not conceptual categories is considered the relevant process.

An important theoretical concept in the perception and representation of categories is that of a perceptual space. A perceptual space is best viewed as an n-dimensional space spanned by psychophysical axes, such as loudness and (perceived) duration. An incoming stimulus is mapped to a point in this space, i.e., receives a value on each of the n dimensions. For example, a well-known way to represent physical colors is the RGB- coding. Each color receives three values ranging from 0 to 255: One value is for red, one is for green and one is for blue. Pure red would, for example, receive (255 0 0), pure blue (0 0 255) and pure green (0 255 0). White would receive (0 0 0) and black (255 255 255). Different combinations of the three dimensions represent all other available colors. With regard to perceptual representations, Shepard (1957; 1987) developed a similar logic. The idea that stimuli could be viewed as points in a space proved to be extremely fruitful. The dissimilarity of stimuli that is so important in their discrimination and hence classification is defined as a function of the *distance* between the two stimuli in such a space. A greater distance between the coordinates of two stimuli in their perceptual space makes them easier to discriminate (Shepard, 1962a, 1962b). Thus, in the (physical) RGB example, it would be much easier to discriminate white (255 255 255)

from yellow (255 255 0) than it would be to discriminate white from ivory (255-255-240).

Tversky (1977) and Tversky and Gati (1982) challenged the concept of a perceptual space when they showed that not all the required underlying assumptions hold in all situations. Tversky (1977) showed that similarity judgments can be asymmetric: North Korea is often judged more similar to China than China to North Korea. This asymmetry is difficult to account for in a metric conception of (dis)similarity based on distance. Furthermore, Tversky & Gati (1982) showed that certain situations violate an important assumption of any distance based space, the triangle inequality. The triangle inequality states that the distance between two points is smaller or equal to the summed distance between those two points and a third point[1]. This holds in physical spaces, but it is not always true of similarity. William James already pointed to violations of the triangle inequality in The *Principles of Psychology* (1890): *The moon is similar to a gas-jet; it is also similar to a football; but a gas-jet and a football are not similar to each other.*

Because of these and other challenges, the concept of a perceptual space has been subject to various modifications. Shepard (1987) pointed out that independent dimensions are orthogonal, but when dimensions interact with one another, the perceptual space becomes oblique. Nosofsky (1986), following Shepard et al. (1960) and Getty, Swets, Swets, & Green (1979) further developed the idea of dimensional weights to allow the perceptual space to expand or shrink depending on the perceptual saliency of each dimension. If these weights may shift across comparisons, violations of the triangle inequality can be accounted for. General Recognition Theory (Ashby & Townsend, 1986; Ashby & Gott, 1988) combines the concept of perceptual space with probability theory. There, the perceptual effect of

1 Mathematically this is expressed as: $d(x,z) \leq d(x,y) + d(y,z)$.

repeated presentations of a stimulus is a probability density function instead of a point (Ashby & Perrin, 1988) and the similarity between two stimuli is based on the overlap of the stimulus' probability density functions instead of on their metric distance (Ashby & Lee, 1992).

There are several possible representations of categories in perceptual space based on the considerations above: e.g., prototype theories, exemplar theories, decision bound theories (Ashby & Maddox, 1993; 2005) and distribution-based theories (Nearey & Hogan, 1986; Nearey, 1997; Smits, Sereno & Jongman, 2006). These theories of categorization all have different accounts of the representation of categories in perceptual space. Prototype models (Rosch, 1973) represent a category by its prototype, typically the mean stimulus. In categorization, a new stimulus is compared to all available prototypes (means) and is assigned to the category with the best matching prototype. Exemplar models, on the other hand, do not use means to represent categories, but instead represent them by storing all the exemplars that have been encountered before (Nosofsky, 1986). When a new stimulus has to be categorized, its similarity to all available exemplars of each category is determined. It then is assigned to the category with which it shares the greatest amount of similarity. Decision bound models (Ashby & Gott, 1988, Maddox & Ashby, 1993), partition perceptual space into response regions whose boundaries are stored in memory. Each region represents a different category. An incoming stimulus will fall on one side of the decision bound and is then assigned to the corresponding category. Thus, decision bound models can be viewed as a multidimensional generalization of signal detection theory. Distributional accounts originated primarily in phonetic research (Nearey & Assman, 1986). As is the case in exemplar theory, categories are not represented by a mean value, but also incorporate

information about the distribution of the stimuli in the category. In contrast with exemplar theory, the distributions are stored in a summarized parametric form in distribution theory, rather than as previously encountered exemplars.

How are categories learned? It is widely assumed that statistical learning lies at the heart of category learning (Diehl, Lotto & Holt, 2004) as well as at the heart of language acquisition (Holt, Lotto & Kluender, 1998). Acquiring categories is then equivalent to recognizing the statistical patterns present in the incoming signals. Repeated exposure to stimuli originating from distinct categories will lead to the formation of "clouds" of points in perceptual space. If, after a period of exposure, several more or less distinct clouds emerge, listeners may start to identify each cloud with a category. Note that trial-by-trial feedback can also be considered as a distributional cue, one that totally correlates with category membership.

Statistically based category learning has been most intensively studied with visually presented stimuli (Ashby & Maddox, 1993; Nosofsky, 1990). Of particular interest for present purposes are the differences between learning a unidimensional and a multidimensional category distinction, and the role of supervision in this learning.

Figure 1.1 illustrates the difference between a unidimensional and a multidimensional categorization problem. All four panels display category structures with variation in two dimensions. However, in two panels the optimal solution is unidimensional, while in the other two panels the solution is two-dimensional. Solving the categorization problem presented in the top left panel in Figure 1.1 requires the use of only dimension 1, whereas the problem presented in the bottom left panel requires dimension 2. In contrast, the categorization problems on the right side of Figure 1.1 require participants to use both dimensions when

making their category judgments. The use of only one dimension would lead to many incorrect categorizations in those cases.

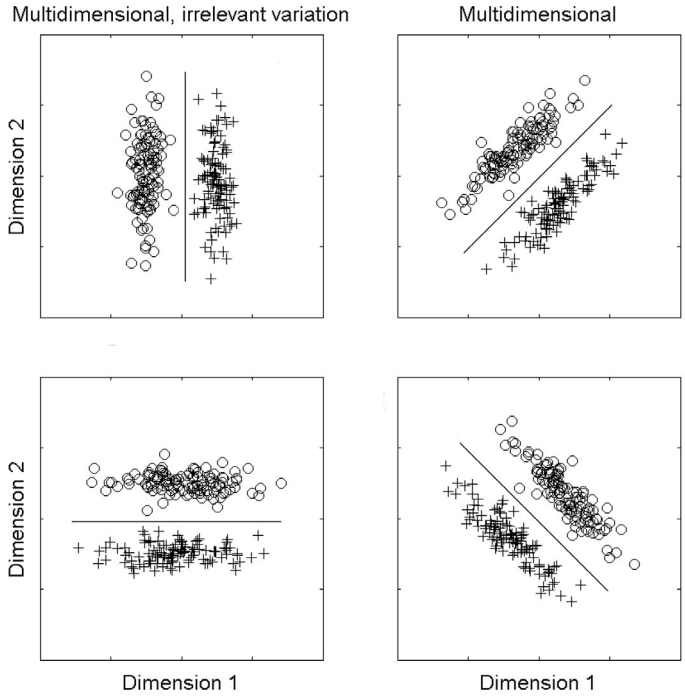

Figure 1.1. Four possible category structures in a two-dimensional perceptual space. Lines represent the optimal solution to the categorization problem.

The distinction between supervised and unsupervised category learning has been extensively studied in adults. Human adults have proven adept at acquiring both unidimensional and multidimensional categories when given regular and immediate feedback about the validity of their judgments on a trial-by-trial basis (Ashby & Alfonso-Reese, 1995; Ashby, Maddox, & Bohil, 2002; Gureckis & Love, 2003). Such feedback is not always required (Fried & Holyoak, 1984; Fiser & Aslin, 2001), however, and is seldom provided by everyday experience. When confronted with complex multidimensionally varying stimuli, learners must rely on the distributional

structure of the objects and events they perceive. When categorization is successful, those stimuli that occupy nearby regions of perceptual space come to be regarded as the same, and as distinct from things that occupy different regions of the same perceptual space. If the correlated structure of category members can be used by the observer, there is a basis for forming a category without external feedback.

Studies of unsupervised category learning have revealed characteristic limits to observers' abilities (Ahn & Medin; 1992; Regehr & Brooks, 1995). Generally, less complex (unidimensional) categories are much easier to learn than complex (multidimensional) ones. Ashby, Queller, and Beretty (1999) showed that participants confronted with a multidimensional categorization problem initially opt for unidimensional solutions (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; the categories differed both in length and orientation (for a graphical illustration, see Figure 1.1). 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.

In line with this result, unsupervised learning of unidimensional rules under conditions where there is irrelevant variation in other dimensions appears to be restricted to situations that are highly structured. Homa & Cultice (1984) had

observers categorize connected dot patterns that differed in their level of distortion of the prototype with and without feedback. Only the condition with low distortion of the dot patterns' prototypes provided enough structure in the stimuli to make unsupervised learning possible. Love (2002) investigated unsupervised learning with the category learning problems constructed by Shepard, Hovland & Jenkins (1961)[2]. Performance was best with Shepard et al.'s type I categorization problem where only one dimension is relevant. With two relevant dimensions (type II), accuracy dropped from 73% correct to 56% correct (Love, 2002).

Most of the evidence supporting the above generalizations derives from experiments testing categories with simple visual stimuli such as lines varying in length and orientation, the size of a circle or the position of dots relative to a mid line (Ashby & Maddox, 1993; Nosofsky, 1990; Feldman, 2000). In these studies, the dimensions of variation are readily identifiable to participants. Artificial categories involving distributions of more complex stimulus patterns whose dimensions of variation are less obvious have not, to our knowledge, been used in unsupervised learning experiments, and, as noted previously, few studies have used these methods to test the learning of auditory categories. Two notable exceptions are Wade and Holt (2005) and Holt and Lotto (2006).

Wade and Holt (2005) had participants play a computer game where sounds originated from different unidimensionally varying categories. The dimensions of variation were the increase and decrease of the spectral frequency at either the onset or the offset of the stimulus. These categories were predictive of the emergence of

2 Shepard et al. constructed six category learning problems with eight stimuli varying in three binary dimensions (for example, round-square, black-white, small-large). The type I problem has only one relevant dimension. In type II, two dimensions are relevant and the other dimension varies irrelevantly. In type VI, all three dimensions are equally relevant and solving the category problem basically means memorizing each categories members. Types III, IV and V are between II and VI in complexity.

different characters in the game. Playing the game became progressively more difficult without paying attention to auditory cues. After 30 minutes of play, participants showed reliable learning at a categorization task showing that participants were able to incidentally (and thus unsupervised) pick up on the statistical information available to them in the auditory input (Wade & Holt, 2005). Holt & Lotto (2006) showed listener biases toward certain dimensions when learning a two-dimensional category distinction. In their experiments, they trained listeners to categorize stimuli differing in two equally salient and equally informative frequency measures (the center frequency and modulation frequency of a sine wave). Despite Holt and Lotto's efforts to equalize the dimensions, listeners displayed an initial preference for one dimension (center frequency). The preference for the center frequency dimension could only be altered by increasing the variance and thus decreasing the informativeness of that dimension. Decreasing the informational weight associated with the preferred dimension was not sufficient to alter the learning strategies of the listeners. Dimensions of equal salience and importance are thus not always considered equal in a categorization task by listeners (Holt & Lotto, 2006).

Acquiring speech categories

The acquisition of the sound categories of a language can occur in two situations: the situation where the infant acquires its first language without any categories being present and the situation of learning a second language while there is already at least one language in place. Infants have to learn to categorize the incoming sounds into the sound categories of their native tongue; learners of a second language face the

problem that the sound categories of the new language can be different from their native language. They have to attempt to integrate the new sound categories into the existing system.

Despite the possible differences between the situation of the infant and the learner of a second language, we hypothesize some of the underlying processes to be the same. After all, the task faced by infants learning a first language and adults learning a second language is the same. Both have to recognize statistical patterns in stimuli that vary on many relevant dimensions. Speech, we argue, is an inherently multidimensional phenomenon. Although there have been significant attempts, notably by Blumstein and Stevens (1979, 1981), to find unidimensional invariants that differentiate between phonetic categories, almost all aspects of the speech signal are now considered relevant for the listener (Diehl & Kluender, 1987). Depending on context and conditions, different parameters of the speech signal come to be the most relevant ones (Cutler & Broersma, 2005). Listeners will generally use all the potentially relevant cues available to them (Diehl & Kluender, 1987; Diehl, Lotto & Holt, 2005).

Thus, the infant's task of acquiring phonetic categories seems hard: the sounds they have to acquire vary on many relevant dimensions, and display considerable overlap between and variability within categories. On top of this, infants are unable to get feedback when trying to learn these categories. Nevertheless, infants are well on their way to learning the phonetic categories of their native language within the first year of life (Jusczyk, 1997). Numerous experiments have demonstrated the ability of infants to discriminate a broad range of speech sound contrasts early in development (in fact, from directly after birth). Over the course of the first year, infants begin to lose the ability to discriminate phonetic contrasts that are not

phonetically relevant in their native language (see Aslin, Jusczyk, & Pisoni, 1998, or Jusczyk, 1997, for reviews). Studies by Werker and colleagues found decrements in non-native phoneme discrimination in English infants at the age of 10-12 months but not at the age of 8-10 months (Pegg & Werker, 1997, Werker & Tees, 1983, Werker & Lalonde, 1988). They concluded that the decline in the ability to discriminate non-native vowels happens in the first year of life and is a function of language-specific experience. Studies investigating non-native vowel discrimination have found decrements in discrimination even earlier in development (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Polka & Werker, 1994). These changes in discrimination ability are seen as adaptive for native language understanding.

It has already been pointed out that corrective feedback of any kind cannot be responsible for infants' perceptual knowledge of their native language, because infants display evidence of this knowledge before they can articulate any words. Moreover, learning the relevant phonetic contrasts on the basis of semantically contrasting minimal pairs (words differing in exactly one phonological feature or segment like /bear/ and /pear/) is also excluded for infants, because their vocabularies contain at best a few meaningful words when they start acquiring phonetic categories (Swingley, 2003). As a result, it is generally assumed that infants acquire their knowledge about phonetic categories by an analysis of the distributional properties of the speech they hear, i.e., through statistical learning. Supporting this notion, experimental studies have shown that infants are extremely sensitive to the statistical properties of incoming speech signals (Kuhl, 2000; Saffran et al., 1996). Simply by being exposed to speech, infants acquire their native-language phonetic categories and lose the ability to discriminate non-native speech contrasts.

Maye, Werker, and Gerken (2002) demonstrated this sort of learning in a laboratory setting in a study of 6- and 8-month old infants. They exposed two groups of infants were exposed to /da/-/ta/ stimuli. One group listened to stimuli in which the *voice onset time* (the dimension differentiating /d/ and /t/) followed a unimodal distribution, encouraging the infants to group the sounds into one category, while the other group listened to stimuli in which the VOT followed a bimodal distribution, encouraging them to group the sounds into two categories. In a preferential looking procedure, infants exposed to a bimodal distribution listened longer to trials with alternating stimuli (two different stimuli) compared to trials with non-alternating stimuli (the same stimulus repeated), while infants exposed to a unimodal distribution did not show this differential looking. The differential effect of the stimulus distributions on the infants' looking behavior shows the sensitivity of infants to the distributional properties of the stimuli.

Adults displayed a similar sensitivity to distributional information in experiments that used similar stimuli (Maye & Gerken, 2000, 2001), which points to similar principles underlying infant and adult categorization (see also Gureckis & Love, 2004). Although Pierrehumbert (2003) doubts the generality of this extremely rapid distributional learning in infants (see also Tyler & Johnson, 2006), there is little doubt that statistical learning underlies the formation of phonetic categories in infants as well as adults.

The literature on adults second language learning (for a review, see Strange, 1995) has demonstrated that it is extremely difficult for adult listeners to master non-native phonetic distinctions at a native level (Burnham, Earnshaw & Clark, 1991). One reason is that the already present native phonological system heavily determines the phonetic category learning problem faced by adults learning a

second language (Cutler & Broersma, 2005). Best's *Perceptual Assimilation Model* (Best, McRoberts & Sithole, 1988; Best, 1995) provides an account of why adult learning of non-native phonetic categories is so difficult. Once the native language categories have been acquired in infancy, the model distinguishes five situations with regard to learning a new phonetic distinction.

The most difficult learning situation described in the Perceptual Assimilation Model is when two distinct non-native phonemes map equally well onto one single native phonetic category. For example, Japanese listeners experience extreme difficulty distinguishing /r/ from /l/ because these phonemes map to a single Japanese phonetic category. It has proven very difficult for Japanese listeners to learn to distinguish these contrasts at a native or near native level (Logan, Lively & Pisoni, 1991; Lively, Pisoni, Yamada, Tokura & Yamada, 1994).

When one non-native category maps reasonably well onto one non-native category and the other non-native category does not, learning depends on the relative goodness of fit of both categories to that native category. A large difference in goodness of fit results in better non-native category learning.

Learning is easier in the two category case, where two non-native phonetic categories map onto two more or less corresponding native phonetic categories. The correspondence between the native and non-native categories does not have to be total; as long as the native and non-native categories map consistently onto each other they can be easily distinguished.

The final two cases represent situations where mapping to the native phonetic system does not happen. Either the non-native phonemes cannot be categorized in the native phonetic system because the non-native phonemes are too far removed from any native category in phonetic perceptual space, or the non-native phonemes

cannot be assimilated into the native phonetic system because they are not perceived as speech. This non-assimilable case is rare. A well-known example is the learning of Zulu clicks by English speakers. Non-native listeners do not consider clicks speech, but find them as easy to discriminate as natives do (Best, McRoberts & Sithole, 1988; Best, McRoberts & Goodell, 2001).

The Perceptual Assimilation Model has received considerable support from studies investigating Japanese native listeners' perceptions of English (Best & Strange, 1992) and English native listeners' perceptions of German (Polka, 1995) and Dutch listeners perceiving English (Broersma, 2005). Experiments reported by Broersma (2005) show that the perceptual effects of listening to non-native phonetic categories can be dependent on phonological rules, something not predicted by the Perceptual Assimilation Model. For example, while Dutch native listeners are perfectly able to distinguish voiced from unvoiced phonemes, this is never necessary in Dutch at the end of words, because of the final devoicing rule in that language. Consequently, Dutch listeners experience difficulty distinguishing English minimal pairs like *bride* and *bright* that differ in final voicing (Broersma, 2005).

Norris, McQueen & Cutler (2003) investigated a category learning strategy available to adults but not to infants. In adults, perceptual adaptation to, for example, a foreign accent can be mediated by lexical support. When a shift in a category boundary between [f] and [s] resulted in perceiving a word instead of a nonword, listeners quickly started to shift this boundary. This perceptual flexibility displayed by adults has been shown to be very talker and context-specific (Eisner & McQueen, 2005), and is stable across at least 12 hours (Eisner & McQueen, 2006).

Thus, although there are lexical as well as statistical sources of information available to adults when adjusting their phonetic categories, their learning

performance with acquiring *new* phonetic categories is not nearly as impressive as that of infants. A longitudinal study trying to train Japanese listeners on the perception of the English /r/-/l/ contrast showed that some improvement is possible, but only after huge amounts of training (Logan, Lively & Pisoni, 1991; Lively, Pisoni, Yamada, Tokura & Yamada, 1994). Moreover, even when non-native categories have been acquired successfully, the contrast between non-native categories is never as quite as sharp as that between native categories (Burnham, Earnshaw & Clark, 1991) and the representations of non-native categories are more talker-specific and contextdependent than in the native language (Lively, Logan & Pisoni, 1993).

This thesis

The experiments presented in this thesis expose adult listeners to categories of sounds, which were either not speech-like (Chapter 2 and 3) or originated from a non-native language (Chapter 4). Because speech categories can differ in more than one dimension, the categorization problems our listeners faced had either one relevant dimension and one irrelevant dimension of variation (a unidimensional categorization problem) or two relevant dimensions of variation (a multidimensional categorization problem). In both cases, the dimensions that varied between the categories were the duration of the sound and the frequency of the spectral peak. These dimensions have been shown to be very important in the perception of vowel sounds (Ainsworth, 1972; Peterson & Barney, 1952).

To construct our stimuli, we defined a two-dimensional perceptual space spanned by perceptual equivalents of duration and formant frequency. The categories were defined as two-dimensional probability density functions in this

space. The distributional characteristics (mean and standard deviation) of the probability density functions governed the relevance of each dimension for making sensible category judgments (see Figure 1.1).

All experiments used a basic procedure with a learning phase and a maintenance phase. In the learning phase, participants listened to stimuli that contained distributional information from the two probability density functions (the categories). Listeners faced the task of partitioning their perceptual space by using one or two dimensions. If participants were to use a unidimensional categorization strategy, all stimuli below a certain criterion value would be assigned to one category and all stimuli above the criterion value to the other category. If a multidimensional categorization strategy were chosen, all stimuli above a criterion value based on a combination of the two dimensions would be assigned to one category and all stimuli below this value would be assigned to the other category (Ashby & Maddox, 1990).

After the training phase, listeners entered a maintenance phase with stimuli that did not contain distributional information (with the exception of the maintenance phase of condition 4 in Chapter 2). With this change in stimulus properties, we wanted to assess listeners' use of each dimension of variation more accurately and to evaluate whether participants would maintain their category identification criteria once the distributional cues to category membership were no longer present in the input. To investigate possible *a priori* categorization tendencies, the experiments with speech stimuli also contained a pretest with stimuli identical to those in the maintenance phase.

Chapters 2 and 3 report experiments studying an analog of the acquisition of a first language, by investigating the learning of nonspeech categories. Chapter 2

investigates supervised learning of nonspeech sounds. Supervision consisted of a visual message indicating whether the response was right or wrong. The results show that it is indeed possible to learn and maintain a unidimensional category distinction. Learning a multidimensional category distinction, in contrast, is much harder, and maintaining this distinction without feedback or distributional cues is very difficult. With distributional cues, however, learning is more easily maintained, illustrating that listeners are sensitive to multidimensional distributional cues in the input.

Chapter 3 deals with unsupervised learning of the same stimulus material as in Chapter 2. The feedback that constituted the supervision was no longer provided. With unidimensional category structures, listeners are sensitive to the distributional information in the input, although performance is not as good as in the supervised case. With a multidimensional problem, the lack of trial-by-trial feedback really hampers listeners, and most of them opt for a unidimensional solution. Chapter 3 also compares listeners' performance in supervised and unsupervised learning.

Chapter 4 is concerned with second language learning. Speakers of Spanish and American English learn to categorize non-native speech using the supervised and unsupervised learning paradigms from Chapters 2 and 3 with unidimensionally and multidimensionality varying stimuli.

Chapter 5 discusses the results obtained and their implications for our knowledge of phonetic category acquisition.

Chapter 2

Supervised learning of phonetic categories[3]

3 Chapter 2 and 3 will be jointly submitted to The Journal of Experimental Psychology as "Supervised and unsupervised learning of auditory categories.

Introduction

Learners of a second language and infants acquiring a first language are faced with the task of learning to categorize the sounds of the language's phonetic system. To succeed in this task, the learner must use phonetic information in the speech signal to determine how many categories there are, and how to categorize additional tokens of sounds as they are heard. Despite a consensus that this process should be conceptualized as a distributional learning problem (e.g., Guenther & Gjaja, 1996; Kuhl et al., 1992; Werker & Yeung, 2005), little is known about the mechanisms by which category learning proceeds, or about what constraints on category learning are present. The experiments presented in this chapter are the first steps in a larger attempt to lay out general principles of auditory category learning, with particular reference to problems posed by phonetic categories.

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

after a period of exposure, distinct clouds emerge, listeners can start to associate each cloud with a different category[4].

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

Infants have been shown to discriminate a wide range of speech-sound contrasts in the first few months of life, but over the course of the first year begin to conflate similar sounds if those sounds are not phonologically contrastive in the infant's native language (see, e.g., Aslin, Pisoni, & Jusczyk, 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 & Tees, 1984) and analogous decrements in non-native vowel perception even earlier (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Polka & Werker, 1994).

Infants' lexical knowledge is almost certainly too meager for language-specific phonological tuning to be driven by semantic contrast in phonologically similar words; thus, infants are generally assumed to learn their language's phonetic categories via a bottom-up distributional analysis of the speech they hear. A

4 This conceptualization of the category learning process was implemented by, among others, Behnke (1998) in an neural network that recognizes patterns and creates a phonetic map. In a similar approach, Kornai (1998) modeled the data of Peterson and Barney (1952) in a neural network.

demonstration of such learning in a laboratory setting was provided in a study of 6- and 8-month-olds by Maye, Werker, and Gerken (2002). In their study two groups of infants were exposed to stimuli on an artificial voice-onset-time (VOT) continuum extending from [da] to unaspirated [ta], a distinction not made in English. One group listened to stimuli in which the VOT followed a unimodal distribution (most sounds were from the middle of the continuum) whereas the other group was presented with stimuli following a bimodal distribution (most sounds were from near the edges). Following this familiarization, infants were given the opportunity to listen to alternating stimulus sets (both of the endpoint stimuli) or non-alternating sets (the same stimulus repeated). Only the infants in the Bimodal familiarization group evinced a preference for alternating over non-alternating stimuli at test, revealing discrimination; infants in the Monomodal group showed no preference. Maye and Gerken (2000, 2001) found a similar sensitivity to distributional characteristics for adults with similar stimuli. However, the generality of this extremely rapid distributional learning is not clear at present (Peperkamp, Pettinato, & Dupoux, 2003; Pierrehumbert, 2003).

In the present chapter we describe experiments in which adult listeners were tested on their ability to learn auditory categories. The categories comprised novel not speech-like sounds with speechlike properties, to simulate processes of phonetic category learning while minimizing effects of native-language phonological knowledge. Listeners' exposure to the category structures was given through experience with category exemplars, in a forced-choice decision task with feedback on each trial.

Our use of artificial categories exemplified by sampling a distribution of variants of category prototypes ultimately descends from the pioneering studies of Attneave

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

We assume that statistical learning lies at the heart of auditory category learning; acquiring auditory categories is equivalent to recognizing the statistical patterns present in the incoming signal. For the purpose of generating experimental stimuli, we specified a psychophysical space spanned by two acoustical dimensions known to be relevant in vowel perception, namely frequency and duration. Categories were defined as two-dimensional probability density functions (pdfs) in this space. Exemplars generated from these functions formed "clouds" in perceptual space. The nature of the pdfs (their means and covariance matrices) governed the relevance of each dimension for making category judgments (see Figure 1.1 in Chapter 1). For example, exposure to the structure in the top left cell in Figure 1.1 in Chapter 1 should encourage subjects to categorize using only dimension 1, whereas exposure to the structure in the bottom left plane should encourage subjects to use only dimension 2. However, exposure to the structures in the right-hand column should encourage the use of both dimensions in categorizing, because the use of only one dimension would lead to many incorrect categorizations. Experiments in visual category learning have shown that subjects initially prefer a unidimensional solution (Feldman, 2000) and only with the help of feedback start using a two dimensional

strategy (Ashby, Alfonso-Reese, Turken, & Waldron, 1998). Ashby et al (1998) distinguish between verbal and procedural-based category learning. In their model, the verbal system has initial priority and this system tries to categorize using a relatively simple (unidimensional) rule (e.g., long sounds in category A, short sounds in category B). Rules that are more complex and more difficult to verbalize like "all long and high frequency sounds go into category A" only enter the verbal system after the unidimensional rules have failed. The other category learning system in their model is an implicit or procedural learning system that does not have such a preference for unidimensional solutions, but learns (much) more slowly.

The notion that learning categories defined over multiple dimensions could be more difficult than learning unidimensional categories may seem counterintuitive. Indeed, category learning is sometimes facilitated by the presence of multiple dimensions of variation. When multiple cues are available to aid in the identification of a category member, or when nominally distinct dimensions' values are interpreted holistically, redundancy gain may be obtained (e.g., Egeth & Mordkoff, 1991; Garner, 1974; Pomerantz & Lockhead, 1991). In addition, the mere presence of correlated attributes among some members of a set of objects can lead observers to form a category that includes those members and excludes the rest—an effect that has been demonstrated even in 10-month-olds (Younger, 1985). However, these advantages of correlations among stimuli depend upon redundancy. Note that in the "diagonal" categories in the right-hand column of Figure 1.1, the value of only one dimension is not a reliable predictor of category membership; good performance requires use of both dimensions. Relative to unidimensional "filtering" tasks (left-hand column), any advantage due to correlations among the dimensions may be outweighed by the fact that listeners must attend to two dimensions rather than one.

The multidimensional-categorization task (sometimes referred to as a condensation task) is more difficult than analogous unidimensional tasks (Posner & Keele, 1970; Gottwald & Garner, 1972).

Distinguishing "diagonal" and non-"diagonal" category distributions presupposes the psychological reality of the axes and a particular interpretation of the axes' orientation. This notion has been studied in attempts to understand the separability or integrality of pairs of dimensions. Broadly speaking, two separable dimensions can be attended to exclusively without mutual interference, while integral dimensions cannot (Garner, 1974). This leads to the prediction that if two category sets defined along separable dimensions are rotated in stimulus space (e.g., converting the left column of Figure 1.1 to the right column), categorization should become substantially more difficult, because observers are deprived of the effective strategy of ignoring the irrelevant dimension (or, conversely, because any tendency to rely on a single dimension leads to many errors). This prediction has been upheld in a number of studies, although the situation is complicated by the fact that classification of dimension pairs as separable or integral is not always maintained consistently over tasks (more thorough discussion of these issues may be found in Grau and Kemler Nelson, 1988; Kemler Nelson, 1993; Melara and Marks, 1990; Shepard, 1991). To anticipate our results, the present experiments reveal a large axis rotation effect, revealing that the speechlike dimensions under study are "psychologically real" in Grau and Kemler Nelson's sense.

Here, learning of multidimensionally varying categories with relevant variation in one dimension was tested in Experiment 1 and learning of multidimensional categories with relevant variation in two dimensions was tested in Experiments 2 and 3. In all experiments, learning was supervised: participants were given feedback

about their category decisions. This contrasts with the situation of the infant in which supervision is absent. We wanted to investigate the role of supervision and make contact with the bulk of the visual perception studies in which such feedback is used. Chapter 3 presents experiments with unsupervised learning and is thus more in line with the infant situation.

All three experiments used the same basic procedure with a learning phase and a maintenance phase. In the learning phase, listeners were presented with stimuli drawn from two probability density functions and received feedback on their category judgments. Listeners were faced with the problem of partitioning the psychophysical space by using a criterion based on one or more dimensions, in the absence of explicit guidance regarding the relevance of the dimensions. Listeners' use of a unidimensional criterion would be reflected in their assignment of all stimuli below a criterion value on that dimension to one category, and all stimuli above it to another. The use of a multidimensional criterion would be reflected by listeners' allowing dimensions to trade off: for example, a low value on one dimension might be compensated by a low value on the other (or a high value on the other, depending on the orientation of the category's "diagonal" in perceptual space). This compensation entails interpretation of one dimension relative to the value of the other in assigning category membership – a process that is a hallmark of speech perception (e.g., Repp, 1982). Solving the categorization problems in Experiment 1 required the use of one dimension, while the categorization problems of Experiments 2 and 3 required the use of both dimensions.

After the learning phase, subjects entered a maintenance phase, intended to characterize their division of psychophysical space. In Experiments 1 and 2 the stimuli for this maintenance phase were drawn from an equidistantly spaced grid

that was intended to "scan" the subjects' psychophysical space in a neutral way, without continued distributional information. In Experiment 3, the maintenance stimuli were identical to those used in the learning phase. In the maintenance phases listeners did not receive trial-by-trial feedback.

The stimuli were inharmonic tone complexes filtered by a single resonance. The two dimensions of variation were the frequency of the spectral peak at which the sound complex was filtered (formant frequency) and the duration of the stimulus (duration). These dimensions are important in the perception of vowel sounds (e.g., Ainsworth, 1972; Peterson & Barney, 1952). We chose to use non-speech sounds as stimuli to prevent subjects' native-language phonetic categories from unduly influencing their category learning (Best & Strange, 1992); however, because these dimensions (or closely related ones) are necessary for speech interpretation, there is no reason to expect that success in the task would require the development of genuinely novel features or stimulus dimensions (see Francis and Nusbaum, 2002, for discussion and evidence bearing on this point for speech sounds, and Schyns, Goldstone, and Thibaut, 1998, regarding feature creation more generally). For example, given that the native language of the participants was Dutch, all subjects were fully accustomed to distinguishing the vowels in words like *maan* ("moon"), *man* ("man"), and *men* ("people"). The first two words' vowels differ primarily in their duration (Nooteboom & Doodeman, 1972), while the last two words' vowels differ in their formant frequencies. Thus, although the inharmonic tone complexes did not sound like spoken words, the dimensions of variation themselves were not new.

Experiment 1

Method

Subjects

Twenty-four subjects (twelve in each condition), all students from the University of Nijmegen (Netherlands), participated in the experiment in return for a small payment. None of the subjects reported any history of hearing problems.

Stimuli

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

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

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

After the base signal was constructed, it was filtered with a single resonance peak, implemented as a second order Infinite Impulse Response (IIR) filter. The filter's bandwidth was 0.2 times that of its resonance frequency. Each sound was

truncated at the desired duration, applying linear onset and offset ramps of 5 ms to avoid the perception of clicks.

In all experiments, the stimuli varied in two dimensions: the frequency of the spectral peak at which the sound complex was filtered (the non-speech analogue of formant frequency) and the duration of the sound. The psychophysical scale commonly accepted for the perception of frequency is the Equivalent Rectangular Bandwidth scale (Glasberg & Moore, 1990). With this scale, physical frequency *f* expressed in Hertz is transformed to "psychological frequency" *e* expressed in ERB units as follows (*f* refers to frequency in Hertz):

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

Psychological duration D (measured in DUR), the psychological counterpart of physical duration in seconds, is converted from stimulus duration according to the following transformation:

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

This transformation was proposed by Smits, Sereno, and Jongman (2006) based on data published by Abel (1972). To ensure that both dimensions would be equally salient and discriminable, they were normalized using their respective just noticeable differences (jnd). The relevant jnd in this frequency region for formant frequency is 0.12 ERB (Kewley-Port & Watson, 1994). For duration, experiments by Smits et al. (2006) and subsequent piloting with multidimensional stimuli varying in duration and frequency indicated that a jnd of 0.25 DUR resulted in a

discriminability comparable to 0.12 ERB. We used these values to equalize the range of variation between the stimulus dimensions, so that the difference between the category means and the in the training distributions and between the highest and the lowest stimulus value in the grid used in the maintenance phase was 20 jnds for both frequency and duration.

Table 2.1.
Distributional characteristics of the stimuli for the two learning conditions (relevant variation in one dimension) of Experiment 1.

	Category A			Category B		
	Means	σ	ϱ	Means	σ	ϱ
Condition 1 (duration relevant)	47.7 D 117 ms 18.80 ERB 1501 Hz	0.65 D 1.07 ms 1.88 ERB 51.3Hz	-0.05	52.53 D 205.0 ms 18.90 ERB 1520 Hz	0.65 D 1.07 ms 1.88 ERB 51.3 Hz	-0.10
Condition 2 (frequency relevant)	50.1 D 149.6 ms 17.6 ERB 1295 Hz	6.45 D 1.91 ms 0.31 ERB 7.76 Hz	0.05	49.73 D 144.5 ms 20.0 ERB 1737 Hz	6.46 D 1.91 ms 0.31 ERB 7.76 Hz	0.10

Table 2.2.
Distributional characteristics of the maintenance phase (equidistantly spaced grid).

	Mean	Min	Max	Step-size
Duration	50.1 DUR 150 ms	47.6 DUR 117 ms	52.6 DUR 193 ms	0.45 D/step
Formant frequency	18.8 ERB 1499 Hz	17.6 ERB 1288 Hz	20.00 ERB 1739 Hz	

Solving the categorization problem in Experiment 1 required the use of only one dimension. The difference between Conditions 1 and 2 was in the relevant dimension of variation. In Condition 1, the stimuli manifested relevant variation in duration and irrelevant variation in formant frequency (see the upper middle panel

of Figure 2.1). In Condition 2, the stimuli manifested relevant variation in formant frequency and irrelevant variation in duration (see the middle panel of Figure 2.1). Table 2.1 shows the perceptual and physical characteristics of the distributions of the learning stimuli of each condition.

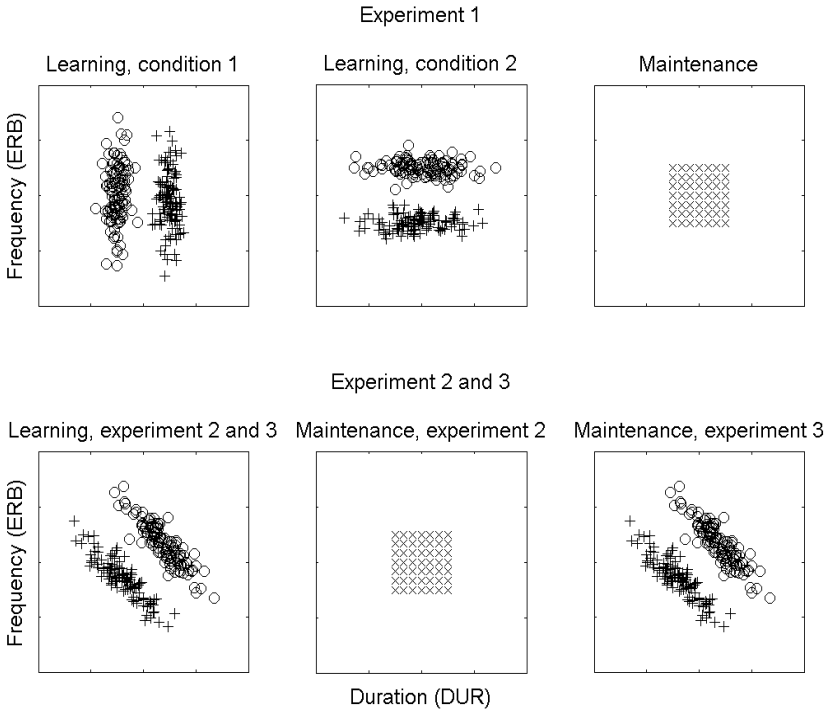

Figure 2.1. Learning conditions of Experiment 1 (upper left and middle panel) and Experiment 2 and 3 (lower left right panel) and test conditions of Experiment 1 (upper right panel), Experiment 2 (lower middle panel) and Experiment 3 (lower right panel).

The maintenance stimuli were the same for both conditions, with items taken from an equidistantly spaced grid (see the upper right panel of Figure 2.1 and Table 2.2). Their values ranged between the mean values of both categories. The test stimuli were intended to "scan" the subjects' psychophysical space in a neutral way, with distributional information no longer present.

Procedure

Subjects were seated in a soundproof booth in front of a computer screen and a two-button response box. In the learning phase, they listened to 448 stimuli (2 categories times 112 stimuli per category times 2 repetitions) through Sennheiser headphones. The stimuli from the two categories were presented in a random order in two blocks separated by a brief rest period. All 112 stimuli from each category were presented once in each block.

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

In the maintenance phase subjects categorized sounds from the test continuum (see the upper rightmost panel of Figure 2), as belonging to group A or B. There were 49 test stimuli that were randomly ordered in four blocks, totaling 196 presentations. Once a participant had selected a category label on a trial, the monitor would display (the Dutch equivalent of) "next" for 700 ms and the next stimulus was played after a 200 ms delay. No feedback was given on maintenance trials.

Results and discussion

The results were analyzed using percentage correct, d' and logistic regression. Both d' and percentage correct are straightforward measures of performance that are easy to interpret. A disadvantage is that they are based on category membership, ignoring

the coordinates of the stimuli in the multidimensional plane and consequently yielding less fine-grained information about participants' strategies. In addition, they cannot be applied to the data of the maintenance phase, because "correctness" of a response does not apply straightforwardly in the region between the trained category exemplars. Logistic regression, on the other hand, is sensitive to the coordinates of the stimuli, and also may be applied to the data of the maintenance phase.

Agresti (1990) argued for logistic regression as the appropriate analysis for categorical response data with continuous stimulus dimensions. In every regression analysis linear and interaction terms can be entered into the analysis. Because in linear regression the interpretation of an interaction term is often problematic it is usually left out. The present results were analyzed both with and without the interaction term. Of the 72 analyses in Experiment 1 (12 subjects times 2 experimental conditions times 3 experimental parts) only 2 had a significant interaction term and of the 72 analyses of Experiment 2 and 3 only 10 had a significant interaction. Furthermore, the fit of the models with interaction term hardly improved compared to those without. Based on these results and the needless complexity of models with an interaction term we present here only the model without the interaction term.

Signal detection analysis

Listeners' performance in Experiment 1 was fairly good. The four bars on the left-hand side of Figure 2.2 show the percentages correct of the first and second part of the learning phase Condition 1 (duration relevant) and 2 (frequency relevant). Recall

that only the data from the learning phase is analyzed because only there can "right" and "wrong" be clearly assigned.

Figure 2.3 shows the same with d' as the dependent measure, where d' = 0 equals chance. In both conditions and both learning phases, percentages correct and d-primes were significantly above chance (all $p < 0.05$) using t-tests with correction for multiple comparisons.

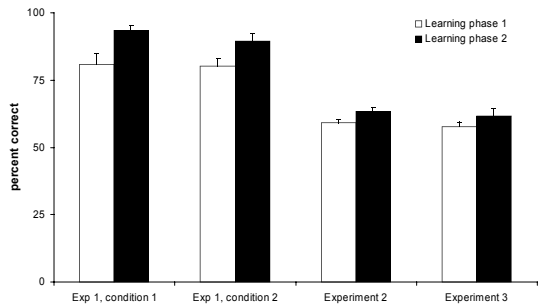

Figure 2.2. Percent correct measures for Experiment 1 to 3.

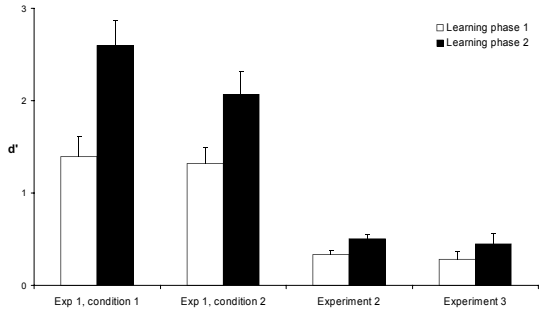

Figure 2.3. d' values for Experiment 1 to 3.

An ANOVA with Part of the Experiment (Learning phase 1 versus 2) as a within-subjects variable and Condition (duration relevant versus frequency relevant) as a between-subjects variable revealed significant improvements in performance from the first phase to the second, whether considering the percent correct measure ($F [1,22] = 7.14$, $p < .05$) or the d' measure ($F [1,22] = 8.31$, $p < .05$). Performance did not vary significantly by Condition (percent-correct and d', $F<1$, ns), nor were there any significant interactions (both $F < 0.1$). See also Table 2.3.

Table 2.3.
Signal detection results; mean percentage correct ("pc") and d', with their standard deviations, for all three experiments.

	Learning phase 1				Learning phase 2			
	pc	σ	d'	σ	pc	σ	d'	σ
Experiment 1, Condition 1	0.81	0.04	1.39	0.21	0.93	0.02	2.59	0.27
Experiment 1, Condition 2	0.80	0.03	1.32	0.17	0.89	0.03	2.07	0,25
Experiment 2	0.59	0.01	0.33	0.05	0.63	0.01	0.50	0.05
Experiment 3	0.58	0.02	0.28	0.08	0.62	0.03	0.45	0.11

Thus, listeners were able to perform the categorization task, and improved from the first phase to the second. In the next section, logistic regression is used to investigate the category learning process in more detail.

Logistic regression

Logistic regression yields, among other things, two β-weights that are similar to the weights in a linear regression. Like the weights in a linear regression the β-weights reflect the influence of the independent variables on the dependent variable. A β-weight of large magnitude indicates a strong influence of the associated dimensions on the dependent variable (the listeners' choice of category). The β-weights were calculated separately for each subject. Table 2.4 and Figure 5 display the mean β-weights for the relevant and irrelevant dimension for the first half of the learning phase ("Learning phase 1"), the second half of the learning phase ("Learning phase 2") and the maintenance phase ("Maintenance phase").

In addition to β-weights, the logistic regression gives significance levels of the hypothesis that each β-weight differs from zero. If a β-weight did not differ significantly from zero at the $p = .05$ level, we concluded that subjects did not make use of that dimension.

Table 2.4 displays the full results of the logistic regression analysis. The columns labeled "Uni" and "Multi" show how many subjects used either one or both dimensions significantly. Numbers of subjects who did not use any dimension significantly are not shown (note that the total number of subjects in each group was always 12).

Table 2.4.
Logistic regression results of Experiment 1 for each condition. Mean β-weights are shown for both dimensions and the number of subjects out of 12 using one (Uni) or both (Multi) dimensions significantly.

	Condition 1				Condition 2			
	Learning phase 1							
	μ(β)	σ(β)	Uni	Multi	μ(β)	σ(β)	Uni	Multi
Relevant	0.65	0.13	10	0	1.37	0.73	11	1
Irrelevant	0.05	0.04	0		0.02	0.03	0	
	Learning phase 2							
	μ(β)	σ(β)	Uni	Multi	μ(β)	σ(β)	Uni	Multi
Relevant	1.50	0.27	11	0	2.28	1.11	11	1
Irrelevant	0.10	0.10	0		0.02	0.04	0	
	Maintenance phase							
	μ(β)	σ(β)	Uni	Multi	μ(β)	σ(β)	Uni	Multi
Relevant	1.54	0.14	12	0	0.20	0.18	9	1
Irrelevant	0.10	0.06	0		0.07	0.06	0	

Table 2.4 and Figures 2.4 and 2.5 confirm that in both conditions subjects learned to use the relevant dimension. Both the mean β-weights and the number of subjects using that dimension were higher than those of the irrelevant dimension. This also shows that subjects were able to ignore irrelevant variation in making their judgments, as the values of the irrelevant dimensions remained close to zero throughout the experiment. The higher mean β-weights and number of listeners using the relevant dimension in Condition 2 compared to Condition 1 suggest that

formant frequency was an easier dimension to learn to attend to. In the maintenance phase, when feedback was no longer given and the stimulus grid was used, listeners persisted in their use of the relevant dimensions. Oddly, however, although formant frequency was easier to learn, it also appeared easier to unlearn, as was evidenced by the large drop in the average β-weight for formant frequency.

To statistically test these effects, we carried out an ANOVA with Part of the experiment (Learning Part 1, Learning Part 2, or Maintenance phase) and Dimension (Relevant versus Irrelevant) as within-subjects variables, and Condition (duration relevant versus formant frequency relevant) as between-subjects variable and the β-weights as dependent measures.

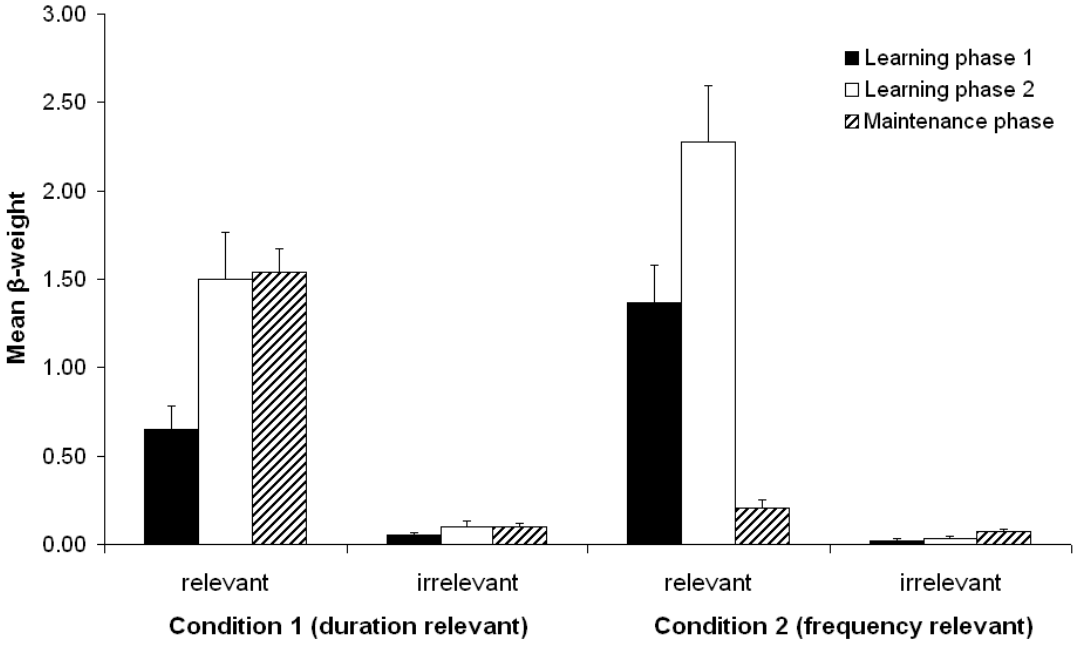

Figure 2.4. Mean β-weights of Experiment 1 for Condition 1 and Condition 2 for the relevant and irrelevant dimensions for each part of the experiment. In Condition 1, duration was the relevant dimension of variation; in Condition 2, formant frequency was relevant. Vertical line segments indicate plus one standard error.

Because of a significant three-way interaction between Dimension, Part of the Experiment and Condition, the results were further analyzed separately for each condition[5]. For Condition 1 (duration relevant), the β-weight for the relevant dimension was higher than that for the irrelevant dimension ($F [1,11] = 61.06$, $p < 0.05$), which confirmed that listeners learned to attend to the relevant dimension. The significant main effect for Part of the Experiment ($F [2,22] = 12.83$, $p < 0.05$) shows that subjects improved over the course of the experiment. The interaction between Part of the Experiment and Dimension ($F [2,22] = 14.40$, $p < 0.05$) indicates that the learning effect depended on whether a dimension was relevant or irrelevant: the effect for Part of the Experiment was present for the relevant dimension ($F [2,22] = 13.78$, $p < 0.05$), but not the irrelevant dimension ($F [2,22] = 1.69$, $p > 0.20$).

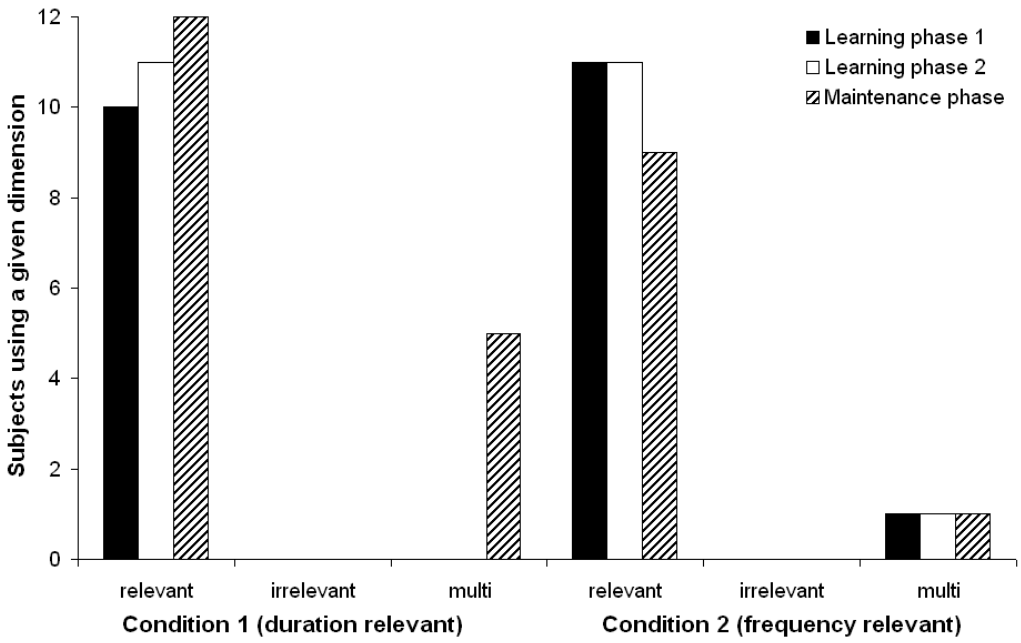

Figure 2.5. Number of subjects using a dimension (duration or formant frequency) significantly in Condition 1 (duration relevant) and Condition 2 (Formant frequency relevant) for all parts of the experiment. Dimensions that have no subjects using them in any condition in any part are not shown.

5 The main effects of Part of the Experiment ($F [1, 22] = 187.98$, $p < 0.05$), Dimension ($F [2, 44] = 13.85$, $p < 0.05$) and Condition ($F [1, 22] = 199.55$, $p < 0.05$) and all interactions were all significant.

In Condition 2, the same main effects and interactions as in Condition 1 were present. The β-weight for the relevant dimension (frequency) was higher than that of the irrelevant dimension ($F [1,11] = 175.04$, $p < 0.05$) and this advantage for the relevant dimension increased during the learning phase (Part of Experiment effect, $F [2,22] = 15.61$, $p < 0.05$). The interaction between Part of the Experiment and Dimension was also present; post-hoc analysis showed a significant effect of Part of the Experiment for the relevant dimension ($F [2,22] = 17.34$, $p < 0.05$), and a much smaller effect for the irrelevant dimension ($F [2,22] = 3.54$, $p < 0.05$). This difference between the conditions reflects the differences in the Maintenance phase. In Condition 1, when duration was the relevant condition, its β-weight remained high in the Maintenance phase and the β-weight for frequency remained small. In Condition 2 however, the β-weight for frequency dropped in the Maintenance phase and that of duration rose. Thus, when presented with an evenly spaced stimulus grid and without feedback, listeners had a tendency to start using duration again even when they had previously correctly used formant frequency.

Differences in the numbers of subjects using a given dimension were statistically evaluated using a binominal test. We wanted to compare the counts of the two relevant dimensions with equal probabilities (0.5) This difference between the counts was significant ($p < 0.05$), confirming that listeners preferred using the relevant dimension over the irrelevant dimension.

Experiment 1 showed a clear learning effect in the learning phase. Subjects learned to attend to the relevant stimulus dimension despite irrelevant variation in the other dimension. A decline in the use of the relevant dimension during the maintenance phase was found for formant frequency, but not for duration. Although listeners who were taught to use formant frequency continued to use it, as evidenced

by the high number of subjects using solely this dimension (10 out of 12 subjects), their performance dropped considerably in the maintenance phase, presumably because feedback fell away and the uniform distribution of the test stimuli no longer supported a distributional distinction between categories. When duration was the relevant dimension, subjects had no problem generalizing the learning to the maintenance phase, as evidenced by the high number of subjects using solely this dimension (12 out of 12) and the consistently high β-weights. The results show that the difficulty in learning to attend to a dimension and maintaining this attention may not be the same for every acoustic dimension. Learning to attend to formant frequency when distributional information was present was easier than doing the same for duration. Maintaining the learned distinction in the absence of feedback and distributional information, on the other hand, was more feasible with duration.

This difference for duration and formant frequency between learning and maintaining a category distinction is surprising given our attempt to equalize the tested dimensions by scaling the variability of the stimuli to empirically determined *jnds*. Apparently, equal just noticeable difference obtained in same/different experiments varying one dimension in a two-dimensional formant frequency x duration space do not guarantee equal categorization behavior. Smits et al. (2006) found a similar difference in their experiments and hypothesized that this may be due to a difference in stimulus dimensions introduced by Stevens and Galanter (1957). Stevens and Galanter argued that dimensions like duration are *prothetic* dimensions, where an increase in value means adding more of the same, while dimensions like formant frequency are *metathetic* dimensions, where an increase does not necessarily mean more of the same. A higher pitch does not mean more frequency, but a longer stimulus duration does mean "more" duration. According to

the model proposed by Smits et al., storing a category representation or comparing a stimulus with a stored category based on a prothetic dimension is noisier than storing a category representation or comparing a stimulus with a stored category based on a metathetic dimension and thus more difficult in the absence of feedback.

Another possibility is that duration and frequency were differentially available to the subjects in these stimuli. That is, to a first approximation the duration of a signal bounded by silence may be measured in a similar way regardless of the spectral characteristics of the signal; but extracting the peak frequency of these tone complexes may have been intrinsically more difficult, or may have profited less from subjects' background experience in processing auditory signals. Although speech makes use of frequency peaks broadly similar to those tested here (and listeners are exquisitely sensitive to variations in these speech features), the present stimuli were not speech signals. If the participants' estimation of frequency was noisier than their estimation of duration, this could have led to their relative disregard for frequency in the maintenance phase (see, for example, Zwicker & Fastl, 1990, pp 265-271). We will return to this issue in Experiments 2 and 3.

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

Experiment 2 addressed learning of multidimensional categories with two relevant dimensions of variation. Instead of what was effectively a unidimensional distinction in Experiment 1, subjects of Experiment 2 had to learn a multidimensional distinction: both duration and formant frequency had to be used in order to obtain a high level of correct responding. This manipulation was

motivated by results from the visual category learning literature, in which learners' performance in unidimensional and multidimensional categorization differ in several ways (e.g., Ahn & Medin, 1992; Ashby, Queller, & Berretty, 1999; Feldman, 2000; Maddox, Ashby, Ing, & Pickering, 2004).

Experiment 2

Method

Subjects

Twelve subjects, students from the University of Nijmegen, participated in the experiment in return for a small payment. None of the subjects had participated in Experiment 1 and none had a history of hearing problems.

Stimuli

In Experiment 2 the main axis of variation of the probability density functions was oriented diagonally (see the lower leftmost panel of Figure 2). To ensure a large enough incentive for participants to actually use both dimensions, we chose the mean and covariance matrices of the two distributions such that using a unidimensional solution to the categorization problem resulted in a much lower optimal percentage of correctly categorized stimuli (70%) than using the optimal two-dimensional solution (100%). Subjects were tested using the same equidistantly spaced grid as in Experiment 1 (see the lower middle panel of Figure 2). Table 2.5 shows the perceptual and physical stimulus characteristics of the learning stimuli for this experiment.

Table 2.5.
Distributional characteristics of the stimuli of the learning condition of Experiment 2 and the learning and maintenance phase of 3 (relevant variation in two dimensions).

Category A			Category B		
Means	σ	ϱ	Means	σ	ϱ
48.38 DUR 126.2 ms 17.79 ERB 1322 Hz	2.80 DUR 1.32 ms 1.34 ERB 35.5 Hz	-0.98	51.66 DUR 175.2 ms 19.70 ERB 1977 Hz	2.82 DUR 1.33 ms 1.33 ERB 35.2 Hz	-0.98

Procedure

The procedure was identical to that in Experiment 1. Note that subjects again did not receive feedback during the maintenance phase.

Results and discussion

Signal detection analysis

As in Experiment 1, the data of the learning phases were analyzed first using the (signal detection theoretic) measures percentage correct and d'. T-tests confirmed that percentage correct exceeded 50, and d' significantly exceeded 0, in both learning phases (all $p < 0.05$, corrected for multiple comparisons). The bars right of the middle in Figure 2.2 show the percentages correct in Experiment 2, while the same is shown for d' in Figure 2.3. Performance was clearly inferior to that of Experiment 1, but performance was above chance.

An ANOVA with Part of the experiment as within-subjects variable confirmed subjects' improvement in the second phase relative to the first, both for the

percentage correct measure ($F[1,11] = 8.78$, $p < 0.05$) and for d' ($F[1,11] = 6.23$, $p < 0.05$).

Logistic regression

As in Experiment 1, the ß-weights as well as the numbers of listeners that used a particular dimension were analyzed using logistic regression. First, we consider the number of subjects who used one or two dimensions above chance levels (see Table 2.6, columns "D", "F", and "Multi").

This analysis illustrated the difficulty of learning a multidimensional category distinction. At most 6 out of 12 subjects learned to use both dimensions during learning and only 4 subjects maintained this ability in the maintenance phase. The increase in number of subjects using both dimensions (from 4 in Learning Part 1 to 6 in Learning Part 2) was due to two subjects who were initially using only duration, but who then learned to also use formant frequency. In the Maintenance phase, subjects generally used duration. To test these effects we compared the use of both dimensions with the use of no dimension at all using a binomial test (with 0.0025 and 0.9975 as a priori probabilities). This showed a significant preference of listeners for the use of both dimensions ($p < 0.05$).

Table 2.6.
Mean values and stand deviations of the polar coordinates ϕ and A of the β weights for duration and formant frequency in the three phases of Experiment 2 and 3, as well as the numbers of subjects using only duration (D), only formant frequency (F) or both (Multi). Subjects using no dimension are not shown.

Experiment 2 (Maintenance with equidistant grid)					Experiment 3 (Maintenance with learning stimuli)				
Learning phase 1									
N = 6					N = 7				
ϕ (σ)	A (σ)	D	F	Multi	ϕ (σ)	A (σ)	D	F	Multi
0.26 (0.12)	0.21 (0.10)	3	0	3	0.30 (0.09)	0.29 (0.14)	2	1	4
Learning phase 2									
N = 8					N = 8				
ϕ (σ)	A (σ)	D	F	Multi	ϕ (σ)	A (σ)	D	F	Multi
0.32 (0.18)	0.34 (0.13)	1	1	6	0.37 (0.03)	0.18 (0.21)	0	1	7
Maintenance phase									
N = 12					N = 8				
ϕ (σ)	A (σ)	D	F	Multi	ϕ (σ)	A (σ)	D	F	Multi
-0.22 (0.31)	0.76 (0.29)	8	0	4	0.24 (0.34)	0.42 (0.18)	0	0	8

The left column of Figure 2.6 presents 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, while the ordinate shows the β-weight for formant frequency (see Nearey, 1997). The data points are divided into four groups: listeners who used both dimensions (identified by asterisks), listeners who used only formant frequency (plus-signs), listeners who used only duration (crosses), and listeners who did not

use any dimension significantly (circles). Optimal performance corresponds to a point in the upper right hand corner of the Figure, at an angle of 45º (when both dimensions are given equal weight) and far away from the origin (reflecting consistent behavior).

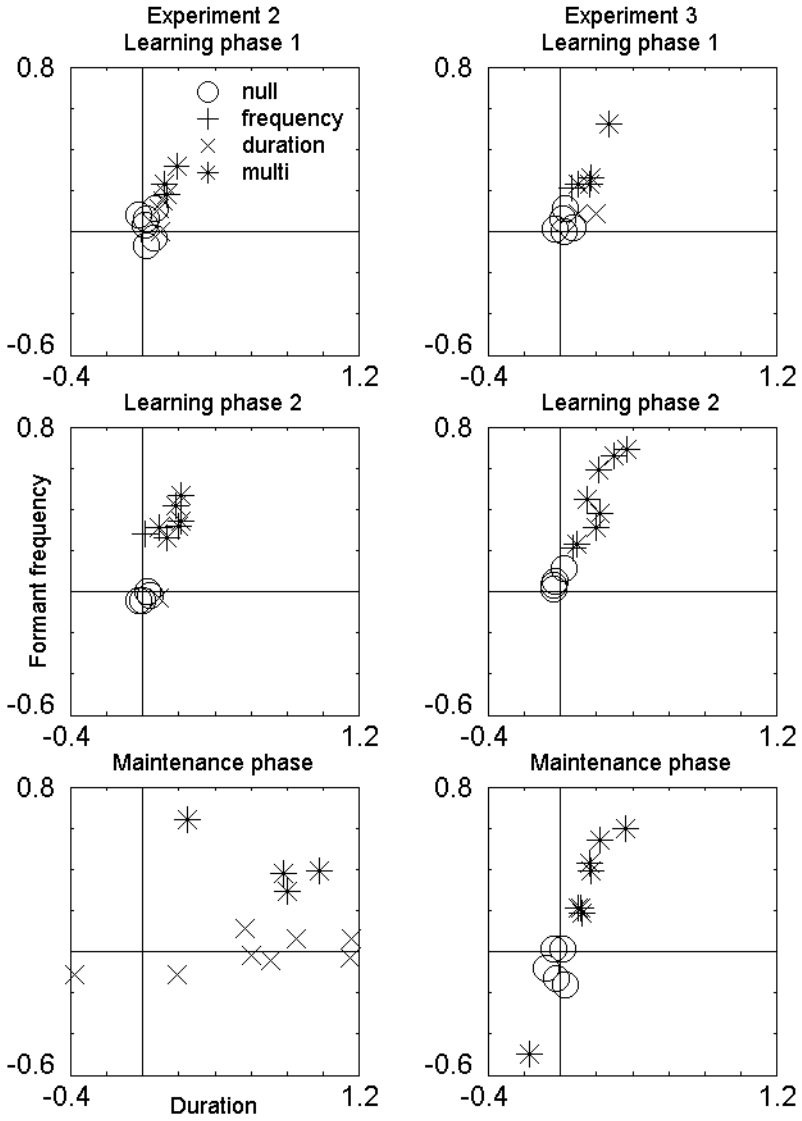

Figure 2.6. Number of subjects using a dimension (duration or formant frequency) significantly in Experiment 2 (two relevant dimensions tested using an equidistantly spaced grid).

The two panels of the left column of Figure 2.6 show performance in the first and second halves of the learning part of Experiment 2. Judging by the number of subjects who used both dimensions in their categorization judgment (the asterisks) a number of listeners picked up on the information provided by the shapes of the categories' distributions and by the feedback. Improvement in the second part is evident in the higher beta values (i.e., asterisks closer to the upper right corner). However, the third panel shows that listeners in the maintenance phase had trouble maintaining their learned categorization strategy (only four asterisks remain) and started using a unidimensional rule with duration as the relevant dimension (the crosses).

Some subjects succeeded in using one or more dimensions above chance levels, and others failed to use any dimensions significantly. For the purpose of comparing the performance of the successful subjects across conditions and experiments, it would be desirable to have a measure of these subjects' central tendency and variability. 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 even though no individuals used both dimensions. These considerations suggest that a measure that integrates performance on both dimensions would be useful.

Here, we derive such a measure by computing the angle formed by the line connecting each subject's Beta weights to the origin, on a graph where the x axis represents duration, and the y axis formant frequency (as in Figure 2.6), and also computing the length of this line. These computations were done first by transforming the Cartesian coordinates of the β-weights for duration and formant

frequency into the polar coordinates ϕ (the angle with the horizontal axis in radians) and A (the distance to the origin) by the following transformations:

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

(5a) $$\phi = \arctan(\beta_{freq} / \beta_{dur}) \; if \; \beta_{dur} \leq 0$$

(5b) $$\phi = \arctan(\beta_{freq} / \beta_{dur}) + \pi \; if \; \beta_{dur} < 0 \; ; \; \phi - 2\pi \; if \; \phi > \pi$$

In our analysis, ϕ ranges between π and $-\pi$ radians. When ϕ equals $\frac{1}{2}\pi$, listeners purely use formant frequency, when ϕ equals 0, listeners use only duration, and when ϕ is close to $\frac{1}{4}\pi$ subjects are in between those two angles and use duration as well as formant frequency. As can be seen from Figure 2.6, listeners who used both dimensions fall in the upper right plane, somewhere between 0 and $\frac{1}{2}\pi$.

The other polar coordinate, A, ranges between zero and 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 Figure 2.6, the listeners that categorized using both dimensions (indicated by the asterisks) are farther removed from the origin, while listeners that do not use any dimension significantly (the circles) are all very close to the origin.

The left hand column of Table 2.6 lists the values of ϕ for each phase of the experiment, considering all subjects who in a given phase used one or more dimensions above chance levels. The mean ϕ of the first part of the learning phase differed significantly from 0 ($t[5] = 5.12$, $p < 0.01$) as well as from $\frac{1}{2}\pi$ ($t[5] = -4.73$, $p < 0.01$). In the second part of the learning phase, the mean ϕ was again significantly different from both 0 ($t[7] = 4.96$, $p < 0.01$) and $\frac{1}{2}\pi$ ($t[7] = -2.88$, $p < 0.05$). In the

maintenance phase, listeners used only duration. The mean ϕ among subjects using any dimension was not significantly different from 0 (t [11], = -0.243, $p > 0.20$), but did differ significantly from $\frac{1}{2}\pi$ (t [11] = -5.850, $p < 0.01$)[6]

An analysis of variance with A as the dependent measure and Part of the Experiment as within-subjects variable revealed a significant effect of Part (F [2,10] = 5.863, $p < 0.05$). Pairwise comparisons showed this effect to be due to a significant difference between the second[7] learning phase and the maintenance phase ($p < 0.05$). Thus, subjects did become more internally consistent in their categorization (higher β weights), but as we have seen, many were becoming consistent in a unidimensional way.

To sum up, learning a multidimensional category distinction was difficult. Where the analysis of percentage correct and d' data did show a learning effect, the values for A and ϕ did not increase significantly from the first learning phase to the second. Moreover, in the maintenance phase both ϕ and A showed that most listeners opted for a unidimensional solution instead of the multidimensional solution suggested by their prior experience. Only half of the subjects used both dimensions significantly during the last learning phase and only four of them retained this ability in the maintenance phase.

Another striking phenomenon is that the advancement of listeners using both dimensions towards the upper right corner is tilted. The line that can be drawn from the origin through the scatter plot is steeper than 45°.This indicates that the mean β-weight for formant frequency is higher that that for duration.

6 Correction for multiple t-tests did not substantially alter the results.

7 The difference between the first learning phase and the maintenance phase is marginally significant at $p < 0.06$.

Most of the subjects who had used both dimensions in their categorizations in phase 2 began to weight duration more heavily in the maintenance phase. Recall that a similar pattern was found between-subjects in the two conditions of Experiment 1: subjects learned to use formant frequency (when it was relevant) more reliably than duration (when it was relevant), but tended to shift toward using duration in the maintenance phase (see Table 2.4).

It is not clear at present why participants were better able to use formant frequency than duration when both feedback and distributional information were present, but appeared to use duration more successfully (Experiment 1) or to a greater degree (Experiment 2) when feedback and distributional information were withheld. As described previously, there are reasons to suppose that duration might be easier to estimate accurately (because it is a prothetic dimension, or because listeners' previous experience makes it easier to measure in these stimuli than formant frequency), but neither suggestion predicts this particular pattern, whose explanation must make reference to differences in the demands of the training and maintenance phases. Although this is not an issue we will resolve in these experiments, Experiment 3 will help clarify the characteristics of the maintenance phase that lead to this result.

There are two possible explanations for participants' change in categorization strategies when they reached the maintenance phase: the absence of feedback in the maintenance phase, and the absence of distributional information (due to the use of an equidistantly spaced grid). Experiment 3 investigated whether the absence of trial-by-trial feedback is in itself enough to disturb the previously learned category boundaries. In Experiment 3 the stimuli of the maintenance phase were no longer taken from the equidistantly spaced grid of Experiments 1 and 2 but were identical

to those in the learning phase. The only remaining difference between the learning and the maintenance phase was the absence of trial-by-trial feedback.

Experiment 3

Method

Subjects

Twelve subjects, students from the University of Nijmegen, participated in the experiment in return for a small payment. None of the subjects had participated in Experiment 1 or 2 and none had a history of hearing problems.

Stimuli

The learning stimuli were identical to those in Experiment 2. The test stimuli were identical to the learning stimuli.

Procedure

The procedure was identical to that in Experiment 3. Again, subjects did not receive feedback during the maintenance phase. In the maintenance phase, like in one of the two learning phases, all stimuli from both categories were presented once in random order.

Results and discussion

Signal detection analysis

In line with Experiments 1 and 2, the data were first analyzed with percentage correct and d' as dependent measures. Both measures differed significantly from chance levels. The rightmost bars of Figure 2.2 and 2.3 show a small difference between the first and second part of the learning phase in the expected direction, though contrary to Experiment 2 this difference between learning phases was not quite statistically significant (ANOVAs, percent correct: $F [1,11] = 2.41$, $p=0.149$; d': $F [1,11] = 3.24$, $p=.099$)[8]

This difference between the learning phases of Experiment 2 and 3 calls for an explanation, because these experiments only differ in their maintenance phases, not in their learning phases. To test whether behavior during the learning phases of Experiment 2 and 3 was significantly different, they were entered together in an ANOVA with Part of the Experiment as within-subjects variable and Experiment as between-subjects variable. This did not yield any significant main effects for Experiment, neither for percentage correct ($F [1,22] = 0.50$, $p > 0.20$) or for d' ($F [1,22] = 0.30$, $p > 0.20$).

Logistic regression

The right-hand column of Figure 2.6 displays the β-weights of each listener in the formant frequency - duration plane. As in the learning phases of Experiment 2, the asterisks show that some listeners learned to use both dimensions in the first learning phase, and that performance improved on this measure in the second

8 An ANOVA examining the learning phases of Experiments 2 and 3 yielded no effects of Experiment nor an Experiment x Phase interaction.

learning phase. This learning was maintained in the maintenance phase of Experiment 3, contrary to the maintenance phase of Experiment 2.

As in Experiment 2, the β-weights were transformed into the polar coordinates ϕ (the angle with the ordinate) and A (the distance to the origin). The right hand side of Table 6 displays the mean values of ϕ and A for each phase of the Experiment. The value for ϕ again lies between 0 and $\frac{1}{2}\pi$, suggesting, on average, the use of both dimensions.

Among those subjects using at least one dimension significantly in each phase of the experiment, mean ϕ differed significantly from 0 ($t[6] = 8.60$, $p < 0.05$) as well as from $\frac{1}{2}\pi$ ($t[6] = -5.60$, $p < 0.05$). This was also true for the second learning phase, where mean ϕ differed from 0 ($t[7] = 35.65$, $p < 0.05$) and from $\frac{1}{2}\pi$ ($t[7] = -0.854$, $p < 0.05$).

Mean ϕ values exceeded $\frac{1}{4}\pi$ (the value that would reflect an unbiased use of duration and formant frequency), indicating more use of the frequency dimension. In the maintenance phase this preference for formant frequency was lost. However, the presence of an outlier in the lower-left quadrant complicates this analysis. With the outlier included, ϕ was marginally significantly different from 0 (duration) ($t[7] = 1.98$, $p < 0.09$) and from $\frac{1}{2}\pi$ (formant frequency) ($t[7] = -2.19$, $p < 0.07$). With the outlier collapsed to the upper right quadrant (on the reasonable assumption that the learner retained his or her knowledge of the categories, but inverted the category assignments), mean ϕ rose from 0.24 to 0.36, reflecting a preference for formant frequency. In this analysis, mean ϕ was significantly different from both 0 ($t[7] = 12.37$, $p < 0.01$) and from $\frac{1}{2}\pi$ ($t[7] = -3.59$, $p < 0.01$) [9]. This is in sharp contradiction with Experiment 2, where consistent maintenance of learning was not found, and in

9 Removing the outlier entirely also yielded a significant difference between mean ϕ for both duration ($t[6] = 40.03$, $p < 0.01$) and formant frequency ($t[6] = -16.01$, $p < 0.01$).

which many subjects shifted to using duration. In Experiment 3, those participants using any dimensions significantly in the maintenance phase all used both dimensions. This difference between two experiments was tested in an ANOVA with Experiment (2 versus 3) as a between-subjects factor and ϕ as the dependent variable. The effect of Experiment was significant ($F[1,17] = 10.24$, $p < .01$).

To test whether listeners became more self-consistent over time, we conducted an ANOVA with the distance parameter A as dependent measure and Part of the experiment as within-subjects variable. This did not yield significant effect of Part of the experiment ($F [2,10] = 0.82$, $p > 0.20$). Pairwise comparisons showed the difference between the first and second learning phases to approach significance ($p < 0.06$), but not the differences between each learning phases and the maintenance phase ($p > 0.20$).

The number of subjects using both dimensions in categorizing the stimuli steadily increased during the experiment from 4 to 7 and remained high in the maintenance phase (8). Compared to the maintenance phase of Experiment 2, the performance of subjects in the maintenance phase of Experiment 3 greatly improved. Analysis with a binomial test comparing the number of subjects using both dimensions significantly with the number of subjects using no dimension at all, showed a significant advantage for the use of both dimensions.($p < 0.01$).

We investigated the effect of the difference between Experiment 2 and 3 (the change in maintenance phase stimuli) by performing an ANOVA on the results of the maintenance phases with Experiment (2 versus 3) as between-subjects factor and A and with ϕ as the dependent variables. For ϕ the analysis (without the outlier) yielded a significant difference between Experiment 2 and 3 ($F [1, 17] = 10.24$, $p < 0.05$) showing the effect of the different maintenance phases. In the maintenance

phase of Experiment 3, ϕ is significantly different from both 0 π ($t [6] = 40.03$, $p < 0.05$) and from ½π ($t [6] = -16.01$, $p < 0.05$) whereas in the maintenance phase of Experiment 2, it only differs significantly from ½π ($t [11] = -5.850$, $p < 0.05$) and not from 0 ($t [11], = -0.243$, $p > 0.20$).

In a final analysis, we compared learning of unidimensional (Experiment 1) and multidimensional (Experiments 2 and 3) categorization problems. Because multidimensional category learning yields two relevant beta's and unidimensional yields one, they are not comparable. Hence, we used the performance measures percentage correct and d' to compare these experiments. An ANOVA with either percentage correct or d' as dependent variable was conducted. Each ANOVA had Part of the Experiment as within-subjects variable and Experiment (unidimensional versus multidimensional, collapsing over Conditions 1 and 2 of Experiment 1) as a between-subjects variable. Significant main effects for percentage correct ($F [1,34] = 6.014$, $p < 0.02$) and for d' ($F [1,34] = 6.278$, $p < 0.02$ were found. Learning a multidimensional distinction was thus significantly more difficult than learning a unidimensional distinction.

General discussion

Listeners provided with trial-by-trial feedback readily learned to differentiate two novel auditory categories that could be distinguished by a single auditory dimension (duration or formant frequency) despite irrelevant variation in the other dimension. Learning a truly multidimensional auditory categorization, on the other hand, proved relatively difficult, even though listeners had at their disposal two sources of

information about the category structure: the distributional characteristics of the category exemplars, and feedback regarding their category judgments.

Participants' success in generalizing to a maintenance phase without supervision depended on whether the relevant dimension was formant frequency or duration, possibly a reflection of processing differences between prothetic or metathetic dimensions (Stevens & Galanter, 1957; Smits et al., 2006) or differences in subjects' ability to extract estimates of duration and of formant frequency from the inharmonic complexes used as stimuli. Performance also depended upon whether the stimuli in the maintenance phase still contained distributional information. If the stimuli in the maintenance phase lacked distributional information, subjects quickly left their learned strategy and reverted to a one-dimensional solution, using the least noisy, i.e. the metathetic, dimension of duration. This result has implications for speech research that uses similar equidistant continua to investigate newly established speech contrasts (Repp & Libermann, 1987), which might be susceptible to rapid degradation resulting from the lack of distributional information at test.

Multidimensional auditory category learning appears to be more difficult than visual category learning, at least based on gross levels of achievement in the present study and analogous visual studies (e.g., Ashby & Maddox, 1993, Nosofsky, 1990). It might be that the stimulus dimensions we chose were particularly difficult ones. Although this possibility is hard to exclude, it seems unlikely given the importance of both frequency and duration for speech

Another difference between the present studies and previous experiments testing visual category learning was the introduction of a maintenance phase without feedback or distributional information. In this maintenance phase, multidimensional category learning performance was notably worse than that in the training phases of

both our and visual category learning experiments. Even very successful unidimensional category learning appeared to be fragile. The lack of trial-by-trial feedback, of distributional information or the amount of training the listeners received to learn the category distinctions are all factors that could be responsible for the difference we observe between performance in the maintenance phases and the learning phases of both visual and auditory category learning.

There are several important issues that remain to be addressed. First, what accounts for the difference between the learning phases and the maintenance phase, especially in multidimensional learning. Second, why do listeners (mostly) prefer duration over formant frequency when left to their own devices? Third, why are there such extensive differences between individuals? Fourth and finally, what do these experiments tell us about infant language learning?

One possible explanation for the difference between the learning phase and the maintenance phase is the absence of feedback in the maintenance phase. When feedback was absent, participants simply "started over", ignoring their previous learning. However, this is unlikely given the lack of such an effect in Experiment 3. A second possibility is that it was the testing of new tokens per se, and not the distributional characteristics of those new tokens, that led to changed performance. This possibility would be more likely if fewer stimuli had been used; however, given that each of the 224 category exemplars was presented only twice during training, it is not plausible to assume that participants had learned to respond to only the set of exemplars themselves; rather, they learned to respond to the categories, with a response strategy generalizable over new exemplars.

We suggest two related accounts of the change. First, in the grid of test stimuli of Experiment 2, the two categories showed no separation; indeed, many of the test

stimuli fell in the region between the trained categories. Such exposure in sufficient quantity should count as evidence to the learner that in fact the two categories are one and the same, for precisely the same reason that distributional learning of categories is possible in the first place. What counts as a "sufficient quantity" should depend on how readily the learner allows new evidence to override earlier, well-supported assumptions. A second factor that may have contributed to the disappearance of multidimensional categorization in the maintenance phase of Experiment 2 is the relatively restricted range of stimulus values in that phase. It is conceivable that the more extreme stimuli of the learning phase "anchored" subjects' memory representations of the dimensions of variation, particularly for formant frequency, and once this variation was reduced, they had more difficulty recovering frequency information from the maintenance stimuli.

The overall pattern of results is consistent with a bias in favor of using duration, except when the distributional characteristics of the presented exemplars contradict duration's diagnostic value.

The definite answer to the question concerning individual differences will be difficult to give. However, there are lots of individual differences in other category learning studies (see, for example, Seger, Poldrack, Prabhakaran, Zhao, Glover, & Gabrieli, 2000). Francis, Baldwin, and Nusbaum (2000) used feedback training to encourage subjects to modify their relative attention to two different cues signaling consonant identity; most of them responded to the training, but several did not. Also in the auditory domain, individual differences have been found in informational masking tasks, in which listeners are required to "listen through" sets of distractor tones in detecting target tones (e.g., Lufti, Kistler, Oh, Wightman, & Callahan, 2003).

Learning a multidimensional category structure is, we argue, a task infants face when acquiring their native phoneme repertoire. Recent studies have suggested that under some circumstances infants can learn unidimensional speech categories without feedback (Maye, Werker, & Gerken, 2002), even when given only 96 stimulus exposures. All current theories of infant phonetic category learning assume that infants can compute categories from phonetic distributions; the Maye et al. (2002) result suggested that this learning might in fact be extremely rapid, helping to account for infants' precocious acquisition of native phonetic categories (e.g., Polka & Werker, 1994). Although there are obviously a number of important differences between the present study and the infant experiments, the current results invite consideration of the possibility that infants' discovery of phonetic categories defined over multiple auditory dimensions is a greater accomplishment than the Maye et al. (2002) results imply. In addition, we suggest that infants, like some of the adults in the present studies, might at first favor unidimensional solutions to multidimensional phonetic problems, or show delayed category learning when the distributional evidence contains trade offs among distinct dimensions. Note that although phonetic cue-trading experiments with infants now have a long history (e.g., Eimas & Miller, 1980), relatively little developmental work has attempted to discover how infants' learning of native-language speech categories is affected by dimensional structure.

Multidimensional learning appeared to be fragile. In the maintenance phase of Experiment 2, we observed that some of the subjects stopped using the multidimensional categorization rule when the distributional information was no longer present. We suggested above that this resulted from the stimulus configuration with which listeners were presented in the maintenance phase. The

use of a grid with equidistantly spaced stimuli to assess the psychophysical space of a listener is a standard technique in the field of phonetics and phonology. The lack of information in the distribution of the stimuli is intended to neutrally probe the subjects' psychophysical space and prevent subjects from changing their categorization tendencies. However, this is not what happened in our experiments; our listeners picked up on the fact that in the maintenance phase the category structure was no longer present, and altered their categorizations. When continuously confronted with stimuli that contained distributional information, their performance level hardly dropped when feedback was discontinued.

Studies of auditory perceptual learning with respect to already known phonetic categories have shown that adult listeners exhibit flexibility in adjusting the boundaries of native language phoneme categories (Eisner & McQueen, 2005; Evans & Iverson, 2004; Francis, Baldwin, & Nusbaum, 2000; Norris et al., 2003; Repp & Liberman, 1987). Such adjustments enable listeners to adapt to new speakers and new dialects.

A positive interpretation of our results would be that our listeners seemed to maintain analogous flexibility towards use of auditory information in the input in our experiments.

The categories which were most speech-like, in that they were defined by truly multidimensional variation, were the hardest for these adult listeners to acquire. By contrast, the categories with only one relevant dimension of variation were well learned despite substantial irrelevant variation in a second dimension.

Our data show that this task is not at all easy for adult listeners, even when they receive feedback. There are several possible explanations for this discrepancy between infant and adult achievement.

First, as Ashby and colleagues (Ashby et al., 1998) have argued, adult participants faced with a new categorization problem will generally start solving the problem in a unidimensional fashion. Only after some training, in which feedback reveals to the learner that the unidimensional approach is not working, will participants switch to a multidimensional strategy (Ashby et al., 1998; Maddox, Ashby, & Waldron, 2002; Maddox, Bohil, & Ing, 2004). In Chapter 3, experiments in which listeners learn the same categories but without explicit feedback that can test this explanation will be presented. A second possibility is that auditory category learning is equally difficult for infants and adults, but that infants simply received much more exposure than the adults had in our experiments. Though short-term modification of infants' speech categories using distributions of unidimensionally-varying exemplars is possible with little training (Maye, Werker, & Gerken, 2002), infants' natural exposure to speech dwarfs our subjects' exposure to the tested categories. On the other hand, categories in natural speech probably exemplify much more variation than our stimuli, because of contextual variability and talker characteristics.

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 comports 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, multidimensional category learning performance collapsed soon after the close of the learning phase.

Chapter 3

Unsupervised learning of phonetic categories

Introduction

The human capacity for resolving the categories of spoken language provides a particularly interesting example of perceptual learning, because the acquisition of language-specific categories begins in infancy (Aslin, Juszcyk, & Pisoni, 1998; Jusczyk, 1997) and because this learning is necessarily unsupervised in nature. This last observation is the starting point of this chapter.

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

Unsupervised category learning studies have revealed characteristic limits in observers' abilities. Ashby, Queller, & Beretty (1999) showed that participants

initially opt for unidimensional solutions (ignoring every dimension of variation but one) and can only be brought to entertain multidimensional solution with the aid of supervision. The studies of Homa and Cultice (1984) and Love (2002) also show the preference for the use of one dimension or category structures with relatively minor prototype distortions.

Most of the evidence supporting these generalizations derives from experiments testing simple visual categories in which the dimensions of variation are readily identifiable to participants (e.g., lines varying in length and orientation; the size of a circle or the horizontal and vertical position of dots relative to a midline). Artificial categories involving distributions of more complex stimulus patterns whose dimensions of variation are less obvious have not, to our knowledge, been used in unsupervised learning experiments, and, as suggested previously, few studies have used these methods to test the learning of auditory categories (cf., Holt & Lotto, 2006).

The literature on visual category formation suggests that in all likelihood, speech sound categories should be extremely difficult to learn. Not only do speech stimuli vary on many relevant dimensions, there is also considerable overlap between categories and variability within categories (e.g., Peterson & Barney, 1952; Hillenbrand, Getty, Clark, & Wheeler, 1995). Yet it is now well-known that infants are well on their way to learning the phonetic categories of their native language within the first year of life. Numerous experiments demonstrate the ability of infants to discriminate a broad range of speech sound contrasts early in development. Over the course of the first year infants begin to lose their ability to discriminate phonetic contrasts that are not phonologically relevant in their native language (see, e.g. Aslin, Jusczyk, & Pisoni, 1998, or Jusczyk, 1997, for reviews). The decrements in

discrimination of non-native consonants (Werker and Tees, 1984) and vowels (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Polka & Werker, 1994) illustrate this point. These changes in discrimination ability are seen as adaptive for native language understanding because the failure to discriminate non-native speech contrasts is taken to imply an improved understanding of the available speech categories in the native language (see Kuhl, et al., 2006, for discussion). In other words, the improved recognition of speech categories of the native language may explain the loss of the infant's ability to discriminate non-native phonemes, possibly because of changes in infants' attention to different phonetic cues. Once two non-native sounds have become part of the same native category, it becomes more difficult to differentiate them from each other and their category co-members (Best, 1995). Within-category discrimination is more difficult than between-category discrimination, because within category sounds are heard as more similar to each other than between category sounds (Cameron Marean, Werner, & Kuhl, 1992; Kuhl, 1985). Given that infants show evidence of perceptual knowledge of their native language before they can articulate any words, corrective feedback cannot be responsible for this learning. Retention of linguistically relevant phonetic contrasts based on semantically contrasting minimal pairs (words phonologically matching in all but one feature or segment) is also excluded for infants whose vocabularies may contain only a few meaningful words (see, e.g., Swingley, 2003, for discussion). As a result, it is generally assumed that infants acquire their knowledge about phonetic categories via an unsupervised bottom-up distributional analysis of the speech they hear (e.g., Pierrehumbert, 2003). This sort of learning was demonstrated in a laboratory setting with infants (Maye, Werker & Gerken, 2002) as well as adults (Maye & Gerken, 2001, 2002[10]). The similar results obtained in the Maye, Werker and

10 For a detailed description of these studies, see Chapters 1 and 2.

Gerken studies for infants and adults points to similar principles underlying infant and adult categorization (see also Gureckis and Love, 2004). Though the generality of this extremely rapid distributional learning in infants and adults has not been determined (Johnson & Tyler, 2006; Pierrehumbert, 2003), there is little doubt that distributional analyses of infant-directed speech provide the foundation of early phonetic category formation.

As in the supervised experiments in Chapter 2, the stimuli were two-dimensional probability density functions in a two-dimensional psychophysical space, as shown in Figure 1.1. The statistical properties of the probability density functions determined the relevance of each dimension for assigning a stimulus to a category. For example, mere exposure to the structure in the top left cell in Figure 1.1 should encourage subjects to categorize using only dimension 1, and exposure to the structure in the bottom left cell should encourage subjects to use only dimension 2. In these "unidimensional" situations, the dimension that does not differentiate the categories is irrelevant to category assignment, although it contributes just as much to the variance of the probability density functions. However, exposure to one of the structures in the right-hand column should encourage listeners to use both dimensions when categorizing the stimuli, because the use of only one dimension would lead to many incorrect categorizations. We assume that recognition of the statistical patterns in the emerging clouds of points in multidimensional space is equivalent to category acquisition. This can be done with feedback (Ashby & Alfonso-Reese, 1995) but learning of perceptual categories without trial-by-trial feedback has also been reported (Fried & Holyoak, 1984; Fiser & Aslin, 2001, Wade & Holt, 2006).

In the experiments presented in this chapter, adult subjects were exposed to categories of non-speech sounds. Although in principle models of adult second language acquisition might best be developed using novel speech categories (such as phonetic categories not present in the language of the participants), it is well known that users of a given language tend to interpret sounds from non-native languages in terms of the perceptual categories of their native language (Best, McRoberts, & Sithole, 1988; Best & Strange, 1992; Flege, 1995; Polivanov, 1931), which complicates efforts to model category acquisition in naïve listeners. Hence, in the present studies the non-speech categories were used here in an attempt to minimize effects of the listeners' native language. Chapter 4 presents experiments that use non-native speech sound in similar supervised and unsupervised learning paradigms.

The stimuli were identical to those used in Chapter 2: inharmonic tone complexes that were filtered by a single resonance. The two dimensions of variation were again the frequency of the spectral peak at which the sound complex was filtered (formant frequency) and the duration of the stimulus (duration).

Both experiments used the same procedure as that presented in Chapter 2 with a learning phase and a maintenance phase. In the learning phase, subjects listened to the stimuli drawn from the two probability density functions. Listeners were faced with the task of partitioning their perceptual space based on one or more dimensions. The use of a unidimensional criterion would be reflected in listeners' assignment of all stimuli below a criterion value to one category and all stimuli above the criterion to another category. A multidimensional strategy would be reflected in listeners' assignment of all stimuli exceeding a criterion value based on a combination of the two dimensions and all stimuli below this value to another (Ashby & Maddox, 1990). In Experiment 1 the categorization problems could be

solved completely (no miscategorized stimuli) by using one dimension, while the problem presented in Experiment 2 required the use of both dimensions. Contrary to the situation in the experiments in Chapter 2, listeners did not receive any trial-by-trial feedback on their categorization in any experiment. After the learning phase, listeners entered the maintenance phase. In the maintenance phase the stimuli were drawn from the same equidistantly spaced grid as in Chapter 2. This change in stimulus properties permitted more accurate assessment of listeners' use of each dimension of variation, and also allowed evaluation of whether participants would maintain their category identification criteria once the distributional cues to category membership were no longer supported in the input.

To investigate the differences between supervised and unsupervised learning, the data from the unsupervised experiments will be compared with their supervised counterparts from Chapter 2.

Experiment 1

Method

Subjects

Twenty-four students from the University of Nijmegen (twelve per condition) participated in the experiment. All subjects were drawn from the Max Planck subject pool and participated in return for a small payment. None of them reported hearing difficulties.

Stimuli

In the preparations for this experiment, a dimension different from duration was at first used in combination with spectral peak frequency, namely the steepness of a rise in fundamental frequency, "sweep". After an extended period of same-different scaling and pilot experiments, the use of this dimension was abandoned. Spectral peak frequency and steepness of the rise in fundamental frequency proved unsuitable for our experimental purposes because their relative salience in the categorization task, which was not predictable from just noticeable differences, almost exclusively determined the categorization behavior of listeners. Appendix A describes the experiments with these stimulus dimensions and the results that led to the present choice of the dimension duration and frequency of the spectral peak.

Thus, the stimuli were identical to those used in Chapter 2: inharmonic sound complexes that varied along the frequency of the spectral peak at which the inharmonic complex was filtered (formant frequency) and the duration of the stimulus (duration). See Table 2.1 and 2.2 in Chapter 2 for detailed stimulus characteristics.

Design

Conditions 1 and 2 differed solely in the relevant dimension of variation. In Condition 1, the stimuli manifested variation in such a way that solving the categorization problem could be done based on duration alone and not on formant frequency (see the leftmost panel of Figure 3.1). In other words, the stimuli in Condition 1 exhibited relevant variation in duration and irrelevant variation in formant frequency. In Condition 2, the stimuli manifested relevant variation in formant frequency and irrelevant variation in duration (see second panel of Figure

3.1) so that solving the categorization problem requires the use of formant frequency only. The maintenance phase of both conditions was identical: listeners categorized stimuli from an equidistant continuum (see the rightmost panel of Figure 3.1) as belonging to either group A or B. This continuum was intended to neutrally "scan" the listeners' perceptual space, as distributional information was no longer present.

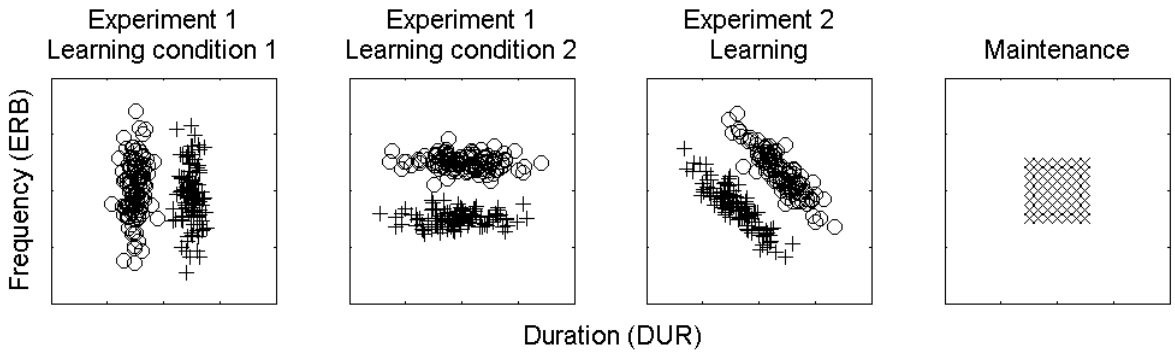

Figure 3.1. Training conditions of Experiment 1 (leftmost and left middle panel) and Experiment 2 (right middle panel) as well as the test condition of both experiments (rightmost panel).

Procedure

The procedure was identical to that of Chapter 2, but without trial-by-trial feedback. The listeners were seated in a soundproof booth in front of a computer screen and a two-button response box. In the training phase, they listened to 448 stimuli (2 categories times 2 repetitions times 112 stimuli per category) through Sennheiser headphones. The stimuli from the two categories were presented in a random order in two sessions separated by a brief rest period. All 112 stimuli from each category were presented once in each session.

The listeners' task was to assign each stimulus to group A or B, using the two-key button box. Once participants had selected a category label on a trial, the monitor would display (the Dutch equivalent of) "next" for 700 ms and the next stimulus was

played after a 200 ms blank screen. No trial-by-trial feedback was provided in the training.

In the maintenance phase the task was to categorize the sounds from the maintenance continuum (see the rightmost panel of Figure 3.1). These consisted of 49 different stimuli randomly presented in four blocks (totaling to 196 stimuli). The maintenance stimuli ranged between the mean values of both categories. No trial-by-trial feedback was given on maintenance trials.

Results and discussion

The results of Experiment 1 were analyzed using percentage correct and the signal detection measure d' as well as measures derived from logistic regression. Percentage correct and d' have the advantage of being easy to interpret measures of overall performance. However, they are based on category membership and not on the coordinates of each individual stimulus in the duration - formant frequency plane. They also cannot be applied to the data of the maintenance phase, as there is no unambiguous criterion for "correctness" of a response there. Logistic regression compensates for these shortcomings, because it is sensitive to the coordinates of the stimuli in the multidimensional plane, which also makes it applicable to the data of the maintenance phase, in contrast to percentage correct and d'.

Logistic regression is the appropriate analysis for categorical response data with continuous stimulus dimensions (Agresti, 1990). As with every regression analysis, logistic regression analysis can deal with linear terms as well as with interaction terms. In logistic regression, this interaction term is difficult to interpret and is therefore usually left out. With our dataset, we ran a logistic regression analysis with

and without the additional interaction term. Of the 72 analyses of Experiment 1 (2 conditions times 3 parts times 12 listeners) only 5 had a significant interaction term. Moreover, the fit of the models with an interaction term was hardly an improvement over those without an interaction term. Due to these observations and the difficulty in interpreting models with an interaction term, we decided to exclude the interaction term in our analysis and to use the linear terms only.

Signal detection analysis

As stated above, percentage correct and d' are easy to interpret summary measures of performance. The two upper rows of Table 3.1 as well as the four left columns of Figures 3.2 and 3.3 list the percentages correct and d's as well as their standard deviations for the two learning phases of both conditions of Experiment 1. The maintenance phase is analyzed in detail using logistic regression.

Table 3.1.
Percent correct and d' for Experiment 1 (Condition 1 and 2) and Experiment 2.

	Learning phase 1				Learning phase 2			
	pc	σ	d'	σ	pc	σ	d'	σ
Experiment 1, Condition 1	0.67	0.17	0.78	0.88	0.76	0.20	1.36	1.16
Experiment 1, Condition 2	0.62	0.13	0.52	0.65	0.71	0.20	0.99	1.04
Experiment 2	0.57	0.05	0.24	0.20	0.59	0.05	0.34	0.19

In both conditions, d' exceeds zero for both learning phases: Condition 1, both phases' $t(11) > 3$, $p < 0.05$; Condition 2, both $t(11) > 2.7$, $p < 0.05$. To test whether percentage correct differed from chance, we first calculated the chance level, which, in an unsupervised learning paradigm, is not equal to 50%. When there is feedback, the mapping of a response to a category can be done a priori and the percentage correct can be calculated accordingly. Without feedback, however, the mapping of

the listener has to be inferred based on her categorization performance. A response most associated with a category is considered to be the one indicating that category. This way, listeners always perform at or above the traditional chance level of 50%. To find the resulting expected value of the assignment of responses to categories they are most associated with, a binomial distribution with this transformed percentage correct was used. This resulted in a test value of 0.5266. With these values, the statistical analysis of percentage correct yield results similar to the analyses of d'. The first learning phase ($t [11] = 2.89$, $p < 0.05$) and second learning phase ($t [11] = 4.04$, $p < 0.05$) of Condition 1 differ significantly from chance and the same held for the first ($t [11] = 2.47$, $p < 0.05$) and second ($t [11] = 3.14$, $p <$) learning phase of Condition 2.

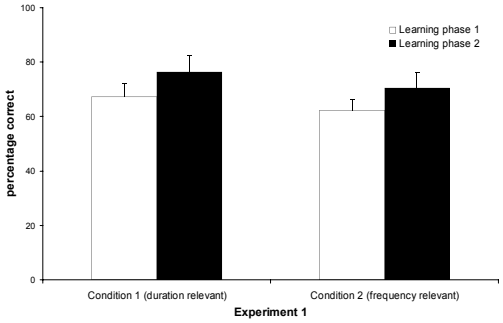

Figure 3.2. Percentages correct in the first and second learning phase of Experiment 1 for Condition 1 (duration relevant) and Condition 2 (formant frequency relevant).

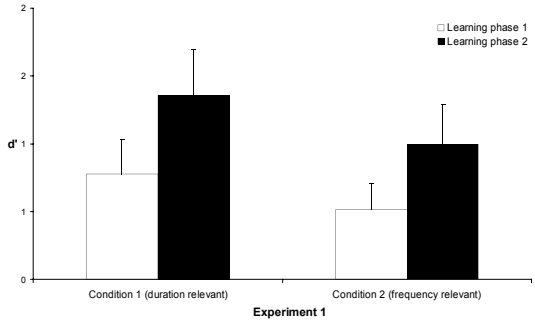

Figure 3.3. d' values in the first and second learning phase of Experiment 1 for Condition 1 (duration relevant) and 2 Condition 2 (formant frequency relevant).

To investigate the effect of learning over time, d' and percentage correct were entered into an ANOVA as dependent variables with Part of the experiment as within-subjects variable and condition as between-subjects variable. For d' there was a significant main effect of Part of the experiment ($F [1,22] = 8.29$, $p < 0.05$) indicating

a higher d' in the second learning phase compared to the d' of the first learning phase. Similar results were, again, found with percentage correct as a dependent variable. The main effect of Part of the experiment was significant ($F [1,22] = 7.14$, $p < 0.05$) showing a significant increase in percentage correct from the first learning phase to the second learning phase.

For both measures, there was no significant interaction between condition and Part of the experiment nor was there a significant main effect of condition. This means that these performance measures are indifferent to whether duration or formant frequency is the relevant dimension.

Logistic regression

Just like a standard linear regression analysis, a logistic regression yields, among other things, β-weights for each independent variable in the equation. These β-weights are comparable to the β-weights in a linear regression in that they modify the influence of the independent variables on the dependent variable (here, the listener's choice of category). A large β-weight indicates that the influence of the independent variable in question is strong, while a small β-weight indicates the opposite. Table 3.2 and Figure 3.5 display the mean β-weights for the relevant and irrelevant dimensions of Condition 1 and Condition 2 for the first part of the learning phase ("Learning phase 1"), the second part of the learning phase ("Learning phase 2") and the maintenance phase ("Maintenance phase").

Table 3.2.
Logistic regression results of Experiment 1 for each condition. Mean β-weights are shown for both dimensions and the number of listeners out of 12 using one (Uni) or both (Multi) dimensions significantly.

	Duration relevant				Frequency relevant			
	Learning phase 1							
	μ(β)	σ(β)	Uni	Multi	μ(β)	σ(β)	Uni	Multi
Relevant	0.38	0.46	5	0	0.47	0.32	6	0
Irrelevant	0.03	0.03	0		0.02	0.01	0	
	Learning phase 2							
	μ(β)	σ(β)	Uni	Multi	μ(β)	σ(β)	Uni	Multi
Relevant	0.75	0.70	7	0	1.03	1.25	6	0
Irrelevant	0.06	0.05	0		0.02	0.01	0	
	Maintenance phase							
	μ(β)	σ(β)	Uni	Multi	μ(β)	σ(β)	Uni	Multi
Relevant	0.98	0.65	9	1	0.75	0.69	5	3
Irrelevant	0.33	0.47	2		0.75	0.74	4	

The logistic regression analysis also indicates whether a β-weight is significant or not. Again, this is identical to the results of a regular linear regression analysis. If a β-weight did not differ from zero at the p = 0,05 level, we concluded that this particular listener did not use that dimensions significantly in categorizing. Table 2.3 lists the β-weights as well as how many listeners use the relevant or irrelevant dimension, or both, significantly. The columns labeled "Uni" and "Multi" convey this information. Listeners who did not use any dimension significantly are not shown, but can be easily calculated, as N is always 12. Figure 3.4 and Figure 3.5 display these results in a bar chart.

Judging by Table 3.2 and Figure 3.4 and 3.5, listeners' performance shows that they learn to use the relevant dimension. The mean β-weight of the relevant dimension is consistently higher than that of the irrelevant dimensions. The same holds for the number of listeners using the relevant dimension compared to those using the irrelevant one. The low mean β-weights for the irrelevant dimensions as well as the small number of listeners using the irrelevant dimension significantly, indicates that listeners not only learned to use the relevant dimension, but also learned to ignore the irrelevant dimension.

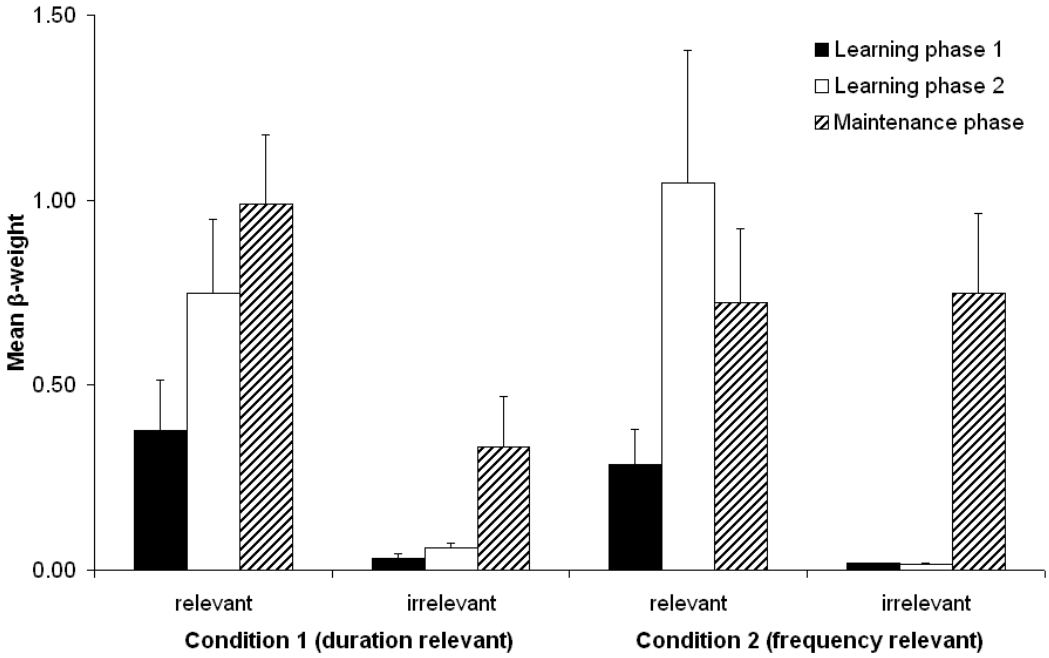

Figure 3.4. Mean β-weights of the relevant and irrelevant dimensions for Condition 1 (duration relevant) and Condition 2 (formant frequency relevant) for each Part of the experiment.

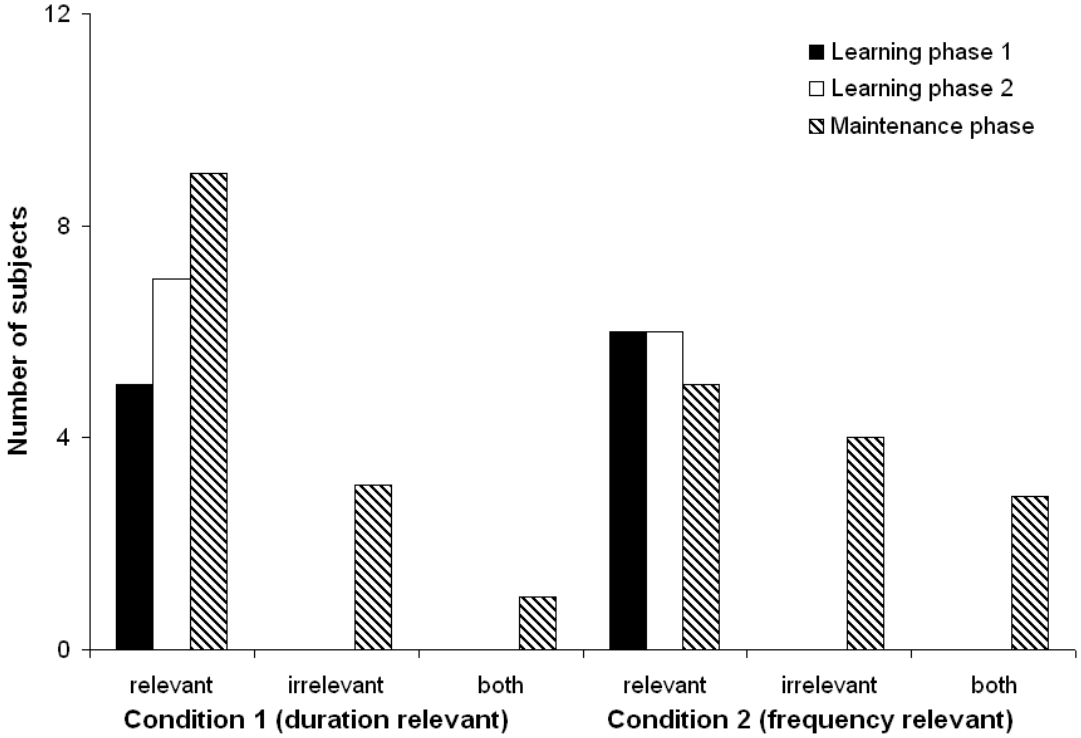

Figure 3.5. The number of subject using the relevant or irrelevant dimension of both for Condition 1 (duration relevant) and Condition 2 (formant frequency relevant) for each Part of the experiment.

The higher β-weights of the relevant dimension in Condition 2 where formant frequency was relevant compared to those of Condition 1 where duration is relevant suggest that formant frequency was somehow the dimension of choice. The same was true of the numbers of listeners using the relevant dimension. More listeners learned to attend to the relevant dimension when formant frequency is relevant, suggesting that this dimension was an easier dimension to learn to attend to. In the maintenance phase trial-by-trial feedback was no longer present and listeners had to categorize stimuli from the equidistant grid. Here, the β-weight and number of listeners using the relevant dimension (formant frequency) showed a drop in Condition 2, but not in Condition 1. Also, in the maintenance phase of Condition 2,

listeners start using the irrelevant dimension duration much more than in the maintenance phase of Condition 1, where formant frequency was the irrelevant dimension. So although formant frequency appeared easier to learn, it is also seemed easier to unlearn.

To test these observations statistically, we carried out an ANOVA with Part of the experiment (Learning phase 1, Learning phase 2, and Maintenance phase) and Dimension (Relevant versus Irrelevant) as within-subjects variables, and Condition (duration relevant versus formant frequency relevant) as between-subjects variable and the β-weights as dependent measures.

The advantage of the relevant over the irrelevant dimension was evidenced by a significant main effect of Dimension ($F [1,22] = 25.17$, $p < 0.05$). The improvement of listeners' categorization judgments over time was reflected in a significant main effect of Part of the experiment ($F [1,22] = 18.79$, $p < 0.05$). Pairwise comparisons showed that each Part of the experiment differed significantly from every other part ($p < 0.05$). There was no significant effect of Condition, nor were there significant interactions. The observed difference between Condition and Part of the experiment, where formant frequency was easier in the learning phase, while there seemed to be a preference for duration in the maintenance phase, did not result in a significant interaction.

We again used a binomial test to assess whether the difference between numbers of listeners using the relevant and irrelevant dimension would differ from chance level (0.5 versus 0.5). In all phases of Condition 1 (duration relevant) the difference between the relevant and irrelevant dimension exceeded chance levels ($p < 0.05$). The same was true in the learning phases of Condition 2 (formant frequency relevant). In the maintenance phase of Condition 2, however, there was no statistically significant

difference between the number of listeners using the relevant (5) versus the irrelevant (4) dimension reflecting the drop in use of the relevant dimension formant frequency.

These data show that listeners can relatively quickly learn a unidimensional category distinction, even in the presence of irrelevant variation in another dimension. The learning effect was most clear in the training phases. In Condition 2, where formant frequency was the relevant dimension, learning was not very robust in the maintenance phase. This difference in dimensions with regard to how easy it was to generalize the learned distinction to the maintenance stimuli was surprising in the light of our effort to equalize the saliency of the dimension in terms of their just noticeable differences. Apparently, equal just noticeable differences obtained in same/different pilot experiments in a two-dimensional formant frequency/Duration space did not lead to equal saliency in a multidimensional categorization task. The same difference between dimensions was found in our previous supervised learning experiments with the same stimuli. There, we pointed to an explanation involving prothetic and metathetic dimensions (Stevens & Galanter, 1957). An increase in value on a prothetic dimension means "more of the same", whereas an increase in value on a metathetic dimension often means a change in quality. A higher pitch does not mean more pitch, whereas a longer duration does mean more duration (Smits, Sereno, & Jongman, 2006). Storing a category representation of a stimulus based on a metathetic dimension is a noisier process than storing or comparing a category representation of a stimulus based on a prothetic dimension. In the absence of feedback, be it trial-by-trial feedback or distributional information, listeners have more difficulty recalling and categorizing a prothetic dimension (Smits et al., 2006).

This brings us to the comparison of supervised and unsupervised learning. Table 3.3 shows the difference scores of both conditions in the supervised and unsupervised learning experiment (supervised minus unsupervised for each performance measure). An overall ANOVA with the signal detection measures (percent correct and d') as dependent variables and the presence or absence of supervision and condition as independent between-subject measures and Part of the experiment as within-subject measure indicated supervised learning to be superior for both percentage correct ($F [1,44] = 20.14$, $p < 0.05$) and d' measures ($F [1, 44] = 18.26$, $p < 0.05$).

Table 3.3.
Difference scores of the unidimensional supervised (Chapter 2) and unsupervised learning (this chapter) experiment. β-weights are shown for both dimensions as well as the signal detection analysis measures for the two learning phases. Positive values indicate an advantage for supervised learning.

	Duration relevant			Frequency relevant		
	Learning phase 1					
	μ(β)	pc	d'	μ(β)	pc	d'
Relevant	0.28	0.14	0.61	1.08	0.18	0.80
Irrelevant	0.02			0.00		
	Learning phase 2					
	μ(β)	pc	d'	μ(β)	pc	d'
Relevant	0.75	0.17	1.23	1.23	0.18	1.08
Irrelevant	0.04			0.02		
	Maintenance phase					
	μ(β)			μ(β)		
Relevant	0.55			0.51		
Irrelevant	0.23			0.67		

The effect of supervision on the β-weights was also investigated. An ANOVA with Part of the experiment (Learning phase 1, Learning phase 2, and Maintenance phase) and Dimension (Relevant versus Irrelevant) as within-subjects variables and Category structure (duration relevant versus formant frequency relevant) and Learning mode (Supervised versus Unsupervised) as between-subjects factor showed a significant advantage for supervised over unsupervised learning ($F [1, 44] = 9.56$, $p < 0.05$). Separate analyses per Category structure were warranted by the significant three-way interaction between Part of the experiment, Learning mode and Category structure. Again, there was an advantage of supervised learning, as evidenced by an effect of Learning mode in Condition 1, when duration was the relevant dimension ($F [1, 22] = 5.07$, $p < 0.05$) as well as in Condition 2, when formant frequency was the relevant dimension ($F [1, 22] = 4.51$, $p < 0.05$). The only difference between Condition 1 and Condition 2 was in the interaction between Learning mode and Part of the experiment. When duration was the relevant dimension, this interaction was not significant, whereas when frequency was the relevant dimension, it was ($F [1, 44] = 17.14$, $p < 0.05$). This interaction reflects the difficulty listeners experience in the maintenance phase of Condition 2 in both Learning modes. With supervised learning, maintaining formant frequency as the relevant dimension was difficult, whereas with unsupervised learning, it was difficult to suppress the irrelevant dimension duration in the maintenance phase.

The results from these two conditions showed that learning of a unidimensional category distinction is possible without the aid of supervision. This is a rather surprising result. Listeners learned to recognize the properties of the probability density functions of the stimuli they listened to, without the aid of trial-by-trial feedback. The extent to which this learning was retained depended largely on which

dimension was relevant. With duration as the relevant dimension, listeners had no problem categorizing the maintenance stimuli according to the learning distributions. When formant frequency was the relevant dimension, listeners were much more sensitive to the distributional properties of the maintenance phase and started using duration more compared to Condition 1. This difference between formant frequency and duration was due, we argue, to noisier encoding of the metathetic dimension (cf. Smits et al., 2006).

The category distinctions in Experiment 1 were all unidimensional in nature. This is in sharp contrast with most speech sounds that have more than one relevant dimension of variation. Lisker (1979) lists seventeen relevant dimensions of variation in his inventarisation of the acoustic features that are involved in the difference between the words *rabid* and *rapid*. Investigating the learning of auditory categories with more than one relevant dimension of variation is quintessential to a better understanding of how people learn to categorize the sounds of their language. Experiment 2 investigated learning of a multidimensional category structure with two relevant dimensions of variation. Listeners had to learn a multidimensional distinction: in order to obtain a high percentage correct, both duration and formant frequency had to be used in the categorization.

Experiment 2

Method

Subjects

Twelve students from the University of Nijmegen participated in return for a small payment. None of them had participated in Experiment 1. All subjects were drawn from the Max Planck subject pool and participated in return for a small payment. None of them reported hearing difficulties.

Stimuli

Table 2.5 (learning phase) and Table 2.2 (maintenance phase) in Chapter 2 list the distributional characteristics of the learning and maintenance stimuli of Experiment 2. Whereas the main axis of variation in Experiment 1 was oriented either horizontally of vertically, in Experiment 2 it was oriented diagonally (see the third panel of Figure 3.1). Listeners were implicitly encouraged to use both dimensions because the mean and covariance matrices we chose resulted in a much lower optimal percentage correct when subjects used a solution with only one dimension (such a solution yielded maximally 70% correct) compared to a solution with two dimensions (which yielded maximally 100% correct). The stimuli in the maintenance phase were identical to those used in Experiment 1 (see the rightmost panel of Figure 3.1).

Procedure

The procedure was identical to that in Experiment 1. Listeners were asked to categorize the stimuli as they saw fit and did not receive trial-by-trial feedback.

Results and discussion

Signal detection analysis

Figures 3.6 and 3.7 display the mean percentage correct and mean d' of Experiment 2. Although performance in terms of these measures obviously was not as good as it was in Experiment 1, the percentage correct of the first learning phases ($t [11] = 2.74$, $p < 0.05$) as well as that of the second learning phase ($t [11] = 3.82$, $p < 0.05$) differed significantly from the appropriate chance level (0.53%). The same was true of listeners' performance in terms of the d' values of the first learning phase ($t [11] = 4.10$, $p < 0.05$) and second learning phase ($t [11] = 6.27$, $p < 0.05$). It should be noted, however, that the d' did not reach the value traditionally associated with good performance in psychophysical experiments (a d' of 1). The distributions of the d' did not overlap completely, but were difficult to separate.

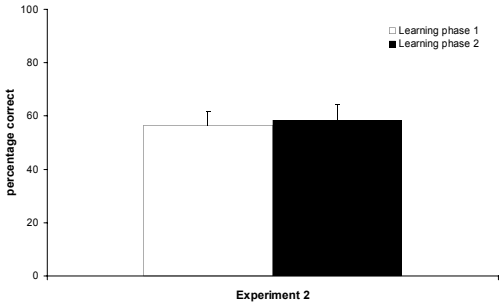

Figure 3.6. Percentage correct for the first and second learning phase of Experiment 2 (multidimensional learning).

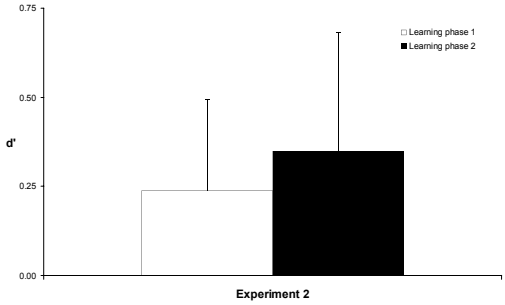

Figure 3.7. d' values for the first and second learning phase of Experiment 2 (multidimensional learning).

The above results already indicate that listeners did notice the distributional information available to them. A significant difference between the first and second phase of the learning phase would be even more evidence of learning. A paired samples t-test did not show a significant difference between the percentages correct of the first and second phase ($t [11] = 1.32$, $p > 0.20$). However, the difference between the d' values of the first and second phase, was marginally significant ($t [11] = 1.93$, $p < 0.08$).

Logistic regression

As in Experiment 1, we conducted a logistic regression analysis with and without an interaction term. Out of 36 regressions, only 4 contained a significant interaction term. Based on this, we decided to use the analysis without the interaction term.

Figure 3.7 and the Table 3.4 display the results of Experiment 2. Figure 3.7 plots the β weights of duration and frequency against one another. Asterisks indicate listeners who used both dimensions, crosses indicate listeners who use duration, pluses indicate listeners who solely use formant frequency and zeros indicate listeners who did not use any dimension significantly at all. In Table 3.4, the columns on the right side display the number of listeners using a given dimension ("D" for the number of listeners using only duration in their categorization, "F" for using only formant frequency and under "Multi" listeners using both dimensions are listed. Table 3.4 does show an increase in the use of two dimensions. Four of our listeners used a multidimensional categorization strategy in the first learning phase and this rose by 4 in the second training phase to 7 in the maintenance phase. There clearly was some sensitivity in our listeners to the distributional properties of the stimuli. Comparing the number of listeners using both dimensions with the numbers

of listeners *not* using two dimensions significantly (hence, all listeners using either only duration, only formant frequency, or no dimension at all) in a binomial test showed that the number of subjects using a multidimensional solution did indeed differ from chance ($p < 0.05$). Due to the relatively small number of subjects and the small odds 0.025 (0.05^2) for the multidimensional solution versus 0.975 ($1-(0.05^2)$)for the other category), expected values sometimes drop below 1, which makes these results difficult to interpret.

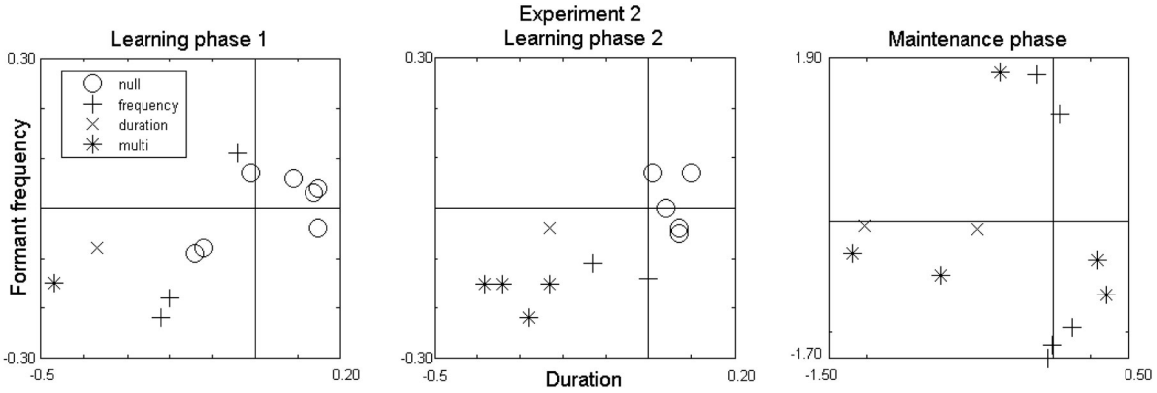

Figure 3.7. The listeners' β-weights associated with the duration and formant-frequency dimensions for the two-dimensional categorization problem of Experiment 2. Plotted in a two-dimensional duration-frequency plane. Asterisks indicate subjects who use both dimensions significantly, crosses indicate listeners who use only duration significantly, plusses indicate listeners who use formant frequency significantly and listeners marked by circles do not use any dimension significantly.

To analyze the multidimensional results presented in Figure 3.7, we transformed the β-weights to polar coordinates. These coordinates represent the angle (Φ) of each individual listeners score with the abscissa and the distance (A) to the origin. When a point in the upper right quadrant is considered, an angle of ¼π indicates a perfectly balanced use of both dimensions, whereas a Φ of ½π indicates the use of only frequency and a Φ of 0 that of only duration (for a detailed description of the logic behind polar coordinates, see Chapter 2). Because listeners' β-weights were often in

the lower left quadrant (which represent a valid multidimensional but mirrored categorization), we recoded the Φ values in that quadrant to Φ values in the upper right quadrant. The left side of Table 3.4 displays these mean polar coordinates for each phase of the experiment.

Table 3.4.
Mean values and stand deviations of the polar coordinates ϕ and A of the β weights for duration and frequency in the three phases of Experiment 2 (multidimensional learning), as well as the numbers of subjects using a only duration (D), only frequency (F) or both (Multi). Subjects using no dimension are not shown.

Learning phase 1 (N = 6) 2				
Φ (σ)	A (σ)	D	F	Multi
0.25 (0.21)	0.31 (0.14)	2	3	1
Learning phase 2 (N = 7)				
Φ (σ)	A (σ)	D	F	Multi
0.20 (0.14)	0.28 (0.10)	1	2	4
Maintenance phase (N = 12)				
Φ (σ)	A (σ)	D	F	Multi
0.20 (0.35)	1.2 (0.42)	5	2	5

We tested whether the values for Φ differed significantly from the two purely unidimensional solutions (represented by Φ's of 0 and ½π). In the first learning phase there was too much variation for mean Φ to significantly differ from either 0 (t [5] = 3.022., *n.s.*) or from ½π (t [5] = -3.27, *n.s.*). In the second learning phase, however, mean Φ differed significantly from both 0 (t [6] = 3.76, $p < 0.05$)[11] or from ½π (t [6] = -5.64, $p < 0.05$). Hence, (some of the) listeners did learn to categorize using both dimensions in the learning phases. In the maintenance phase, the mean Φ differed significantly from ½π (t [11] = -2.95, $p < 0.05$) but not from 0 (t [11] = 1.99,

11 All results incorporate adjustments for multiple comparisons.

n.s.), reflecting a similar preference of the listeners for duration as was found in the maintenance phase of Experiment 1.

To investigate whether the categorizations of the listeners got more consistent over time, the polar coordinate A (reflecting the distance to the origin) is an appropriate measure. An ANOVA with A as the dependent variable and Part of the experiment as a within-subjects variable did not reveal a significant effect of Part of the experiment ($F [1,22] = 1,68$, *n.s.*).

Although not all measures reflected multidimensional learning in Experiment 2, listeners were shown to be sensitive to the distributional information in the stimuli, both in the signal detection theoretic measures, the mean β-weights (as expressed in Φ) and in the numbers of subjects using a dimension. Listeners do perform better with unidimensional category learning problems. An ANOVA with percentage correct and d' as dependent measures, Part of the experiment as within-subjects measure and Orientation (unidimensional versus multidimensional) as between-subjects measure showed a significant effect of Orientation for both percentage correct ($F [1,34] = 6.01$, $p < 0.05$) and d' ($F [1,34] = 6.29$, $p < 0.05$).

Table 3.5.
Difference scores of the multidimensional supervised (Chapter 2) and unsupervised (this chapter) learning experiments. Signal detection analysis measures are shown for the two learning phases and A is shown for all three phases of the experiment. Positive values indicate an advantage for supervised learning.

	pc	d'	A
Learning phase 1	0.02	0.09	-0.10
Learning phase 2	0.02	0.16	0.06
Maintenance phase			-0.44

Finally, multidimensional unsupervised learning was compared with multidimensional supervised learning. Table 3.5 lists the difference scores for the measures that can be compared; percentage correct and d' from the signal detection theoretic analysis and the consistency measure A from the logistic regression analysis. With percentage correct as dependent measure, there was a significant advantage for supervised learning in an ANOVA with Part of the experiment as within-subjects variable and Experiment (Supervised learning versus Unsupervised learning) as between-subject variable ($F [1,22] = 4.98$, $p < 0.05$). For d' no such effect was found ($F [1,22] = 3.55$, $p < 0.07$). A similar ANOVA with the consistency measure A as dependent measure also did not reveal a difference between supervised and unsupervised learning ($F [1,22] = 1.50$, *n.s*).
In summary, Experiment 2 showed it to be possible, but much harder to benefit from distributional information when learning a multidimensional category distinction.

General discussion

The results from both Experiment 1 and Experiment 2 make it clear that unsupervised learning of auditory multidimensional categories is feasible. There were important differences between the learning of unidimensional category distinctions and multidimensional category distinctions as well as between supervised and unsupervised learning.

When there were two relevant dimensions of variation, learning to use both dimensions to correctly categorize the stimuli was much more difficult, but there was not as much difference between supervised and unsupervised learning as was found for unidimensional category learning problems. Listeners were clearly

sensitive to the distributional information present in the stimuli, but not all reached a suitable categorization strategy during the 440 training stimuli. It might be that there were not enough trials to show a larger learning effect, but the absence of a difference between Learning phase 1 and Learning phase 2 suggests either that learning is very slow or that our listeners were already at ceiling. Listeners had a preference for duration over formant frequency when they incorrectly chose a unidimensional solution.

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

In both Experiment 1 and Experiment 2 there was a differential effect of dimension, particularly in the test phase. There, duration seemed to be the dimension of choice. In the absence of distributional cues, subjects were more prone to use duration than formant frequency in their categorization. In the results and discussion section of Experiment 1 we hinted at an explanation for this preference in terms of Stevens and Galanter's (1957) distinction between prothetic and metathetic

dimensions. Smits et al. (2006) argue that the storing and representation of metathetic dimensions as formant frequency is noisier than that of a prothetic dimension like duration. Another possible explanation is to extend Ashby et al.'s (1999) distinction between rules that are easy to verbalize and rules that are hard to verbalize. Especially in Experiment 2, numerous participants reported being at a loss in the test phase and opting for the duration distinction because it was easier to distinguish the sounds based on duration. Further, when asked for the two dimensions of variation, most subjects find it harder to describe the timbre dimension compared with the durational dimension. Formulating a rule in the test phase would accordingly be easier with duration as the relevant dimension. Deciding between these two explanations would require an experiment with two dimensions that are similar in terms of Stevens and Galanter's prothetic/metathetic distinction or that are similar in terms of verbalizability. It is not self-evident, however, how to find a good measure of how easy it is to use a certain dimension to verbalize a rule.

Compared to visual category learning of similar and even more complex category structures (Ashby & Waldron, 1999s) auditory category learning appears to be even more difficult. This could be due to an unlucky selection of particularly difficult stimulus dimensions, although it seems unlikely given the importance of both formant frequency and duration in the perception of speech.

Another important difference of our approach with visual category learning is the unidimensional grid that listeners had to categorize in the test phase. Although performance with auditory categories in the multidimensional category learning experiment was also below expectations based on visual category learning results in the training phase, the test phase yielded the most surprising declines in

performance. Even very successful unidimensional category learning appeared to be fragile in the test phase. The degree of success in generalizing to a test phase without distributional information may depend on whether the relevant dimension is prothetic or metathetic. Confronted with a unidimensional grid, subjects quickly left their learned strategy and reverted to a one-dimensional solution, using the least noisy, i.e. the prothetic, dimension. The use of a test grid with equidistant stimuli is a well-known technique in auditory categorization research, the absence of distributional information is intended to neutrally probe the subjects' psychophysical space and prevent them from changing their newly acquired categorization tendencies. This was not what happened here. Our listeners apparently noticed the change in the distribution of the stimuli in the test phase and altered their categorization behavior to reflect this change. We know of no studies in the field of visual category learning that use a procedure with a training phase where distributional information is present and a test phase where it is absent. These discrepancies warrant further research into the robustness of visual (and auditory) category learning.

The comparison of the supervised and unsupervised learning experiments showed an overall advantage for supervised learning. This was especially clear in the unidimensional learning experiments. There, supervision helped suppress the tendency to use the irrelevant dimension in the test phase. Performance in unsupervised learning of a unidimensional category structure was still surprisingly good, considering that listeners' only source of information was the distribution of the stimuli in perceptual space. When learning of a multidimensional category structure is concerned, the large advantage for supervised learning that was found for unidimensional learning was not present for multidimensional learning. There

was a small advantage for unsupervised learning in the test phase, which might have been due to the similar procedure for the training and the test phase in the case of unsupervised learning. With supervised learning, subjects were faced with the sudden withdrawal of trial-by-trial feedback in the test phase, whereas this was not the case in the unsupervised learning experiments.

Love (2002) compares supervised and unsupervised learning and concludes that unsupervised learning is multifaceted. The variety studied in this paper is best described as intentional unsupervised learning, because listeners were aware of the goal of the experiments. Intentional unsupervised learning is not qualitatively different from supervised learning according to Love. Although our results are not particularly suitable to test this conjecture, there does not seem to be a large *qualitative* difference between our unsupervised and supervised learning results. Our data showed a quantitative difference between supervised and unsupervised learning of unidimensional category structures with better performance with supervised learning, but did not support a difference between the supervised and unsupervised learning of multidimensional category structures.

Learning to categorize auditory stimuli with more than one relevant dimension of variation is, we think, the task infants (and learners of a second language) face when they acquire the sounds of their native language. We have also argued that this learning process is almost certainly unsupervised in nature. However, our data show a large discrepancy between adult and infant achievement. Learning to categorize multidimensional acoustic categories is not at all an easy task for adult listeners, whereas infants all succeed seemingly effortlessly. This discrepancy between our findings and the results from infant research can be explained in several ways.

First, both Ashby et al. (1999) and Love (2002) have argued that subjects will initially opt for a unidimensional solution when they are faced with a new categorization problem. Only when there is sufficient negative feedback will they switch to a multidimensional strategy. Most studies construe this negative feedback as trial-by-trial feedback (Ashby et. al, 1998, Maddox, Ashby & Waldron, 2002, Maddox, Bohil, & Ing, 2003). However, our previous experiments with supervised learning showed similar poor learning of multidimensional categories. Apparently, learning multidimensional auditory categories is not as much influenced by supervision as learning multidimensional visual categories. In the approach of Gureckis & Love (2003) trial-by-trial feedback is not necessary. A surprising event will also change the categorization behavior of the model. Although our listeners clearly were sensitive to the distributional information in the stimuli, the discrepancy between their categorizations and the probability density functions may not have been surprising enough to switch to a multidimensional rule.

A second explanation is that infants receive much more exposure than adults did in our experiments. Though Maye, Werker, and Gerken (2002) have shown that short-term modification of infants' speech categories is possible with very little training, it remains the case that infants' day-to-day exposure to speech dwarfs the 440 stimuli our participants listened to. The everyday speech input infants receive, on the other hand, is much more complex in terms of contextual variability and talker characteristics than our stimuli. Hence, it is difficult to compare the relative difficulties of the learning task faced by infants and the one faced by our listeners. The third possibility is that the difference in learning capacities between infants and adults is simply greater than we thought it was.

Surprisingly little is actually known about the learning of auditory categories in infancy, or even in adulthood. The current experimental evidence about how infants learn phonetic categories points to some sort of distributional learning account in which infants perform a statistical analysis over large numbers of speech sounds and eventually cluster these together in clouds of points.

The experiments presented here studied unsupervised learning of auditory categories by combining techniques borrowed from categorization research in the visual modality with procedures from phonetics and phonology. Listeners were presented with extensive exposure to distributionally defined categories. When faced with a truly speech like categorization problem (a category distinction based on the integration of two dimensions), performance was low and collapsed quickly in the test phase, even though the dimensions spanned a range of 20 just noticeable differences and a unidimensional solution led to 30% more incorrect categorizations. The poor learning in the training phase awaits explanation, whereas the decline in the test phase is most likely due to the flexibility of listeners when confronted with previously unencountered distributional properties. Several studies of perceptual learning of speech have shown the same flexibility of listeners in adjusting the boundaries of the native language phoneme categories (Eisner & McQueen, 2005, 2006; Evans & Iverson, 2004; Norris et al., 2003; Repp & Libermann, 1987). Such quick adjustments enable listeners to adjust to dialectical variation as well as to speaker variability. The listeners in our experiments showed similar flexibility towards auditory distributional variation in their input.

These experiments show that listeners perform well with unsupervised learning of unidimensional non-speech auditory categories despite another irrelevant dimension of variation. This learning is fragile judging by the change in

categorization behavior of listeners when confronted with stimuli without distributional information. Multidimensional learning of multidimensionally defined category structures is possible but difficult and even more fragile than multidimensional learning, despite the fact that these categories are more similar to real speech.

Chapter 4

Supervised and unsupervised learning of speech categories

Introduction

Those who have tried to learn sounds of a foreign language as an adult have undoubtedly sometimes been bewildered by their own inability to grasp a distinction between two non-native phonetic categories. A distinction so fundamental and apparently easy that all users of the foreign language in question, from the oldest adult to the youngest child, take it for granted. What are the processes behind the process of acquiring the sounds of a second language? This chapter tries to investigate a number of processes involved; the role played by the phonology of the first language, the role of the distributional properties of the phonetic categories, and the role of supervision in phonetic category learning.

The literature on the acquisition of non-native phonetic distinctions (for a review, see Strange, 1995) has shown that it is extremely difficult for adults to learn a non-native category distinction, especially at the native or near native level (Burnham, Earnshaw and Clark, 1991). One important reason for the difficulty adults experience in learning a second language is the interference of the native phonology that is already present (Cutler & Broersma, 2005; Best & Tyler, 2006). The native phonological system determines to a great extent how speech sounds are perceived and is thus responsible for the difficulties that arise in distinguishing two non-native speech sounds. A model that describes the various situations encountered in

learning of second language is the *Perceptual Assimilation Model* (PAM) by Best (1995; Best, McRoberts, & Sithole, 1988).

The Perceptual Assimilation Model assumes an adult-like native phonology and distinguishes three options: non-native speech sounds are categorized within the phonological system of the first language, are left uncategorized but still perceived within the native speech system or, a rare case, are not assimilated and are thus not considered to be speech. These three options branch into five situations for the non-native listener. First, when two non-native speech categories are categorized within the native phonological system, the two non-native phonemes might (imperfectly) map to two native phonetic categories (a situation labeled the Two Category case in the PAM). Discriminating the two non-native sounds is easy in this case. The native and non-native categories do not have to be identical, as long as there is a sufficiently consistent mapping between the two native and non-native categories, they can be easily distinguished.

Second, when both categories are categorized within the native phonological system but map onto the same native phoneme (the Single Category case), discrimination is very difficult. A well-known example is the extreme difficulty Japanese listeners experience in distinguishing /r/ from /l/ because these two non-native phonemes map to a single native Japanese phonetic category. Even after extensive training, discriminating these non-native categories is extremely difficult for Japanese listeners (Logan, Lively, & Pisoni, 1991; Lively, Pisoni, Yamada, Tokura & Yamada, 1994).

Third, both categories can be categorized within the native phonological system but one category is mapped to a native phonetic category better than the other category. In this case (The Category Goodness case), non-native category learning

depends on the relative goodness of fit of both non-native categories to the native category. If the difference in fit is large, discrimination and non-native category learning become easier. The categorization of Hindi stops by English listeners is an example of this case. The dental stop matches well its English counterpart, while the retroflex stop is a very poor match. Consequently, the contrast between Hindi dental and retroflex stops is in principle learnable for the English listener.

The fourth case (called the Uncategorized case) is when the non-native speech categories are left uncategorized (i.e., they are not mapped to native speech categories), but are still incorporated into the native phonological system (i.e., they are considered speech). This happens when there are no native phonetic categories that are sufficiently similar to the non-native ones to make mapping possible. The distance in phonetic space between the non-native phonemes and the nearest native phonemes is too large for the native phonemes to successfully assimilate the non-native phonemes. According to Best and Tyler (2006) either one non-native category or both could be left uncategorized. When only one category is left uncategorized, discrimination can be very good because one non-native category is mapped to a native one and the other is not. When both non-native categories are left uncategorized, discrimination is poor or reasonable, depending on the distance of both non-native categories to the closest native phoneme categories.

The fifth and final case is when the non-native phonetic categories are not mapped onto the native phonological system and are thus not considered speech by the non-native listener. In this infrequent case, category discrimination is good to excellent For example, Zulu clicks are usually not considered speech by non-native ears, but non-native listeners discriminate them as well as native listeners (Best, McRoberts & Sithole, 1988).

The Perceptual Assimilation Model has received considerable support from various studies investigating the perception of non-native speech sounds. For example, the perceptions of native Japanese listeners of English (Best & Strange, 1992) of native English listeners of German (Polka, 1995) and that of native Dutch listeners listening to English (Broersma, 2002; 2005) conform to the predictions of the Perceptual Assimilation Model. However, Broersma also showed that the phonological rules of the native language can alter the perception of non-native phonemes. Dutch listeners experience difficulty in distinguishing English minimal pairs that differ in final voicing. In Dutch, the voicing distinction is never relevant at the end of words because of the final devoicing rule in that language (Booij, 1995). Dutch listeners consequently have trouble distinguishing words like *peas* and *peace*. In an attempt to account for the perceptions of listeners that are not naive anymore but have mastered some of the sounds of a second language within the Perceptual Assimilation Model, Best and Tyler (2006) state that perception of non-native phonemes is not only determined by the native phonology but also by phonotactic biases, coarticulatory patterns and allophonic variation.

The Perceptual Assimilation Model shows the importance of the native phonology in learning new phonetic categories. Another important factor in the acquisition process are the distributional properties of the stimuli. The effects of this variation have been extensively studied in visual category learning (Ashby & Maddox, 1993; Nosofsky, 1990). We frame the learning of phonetic categories in a way similar to these studies. There perceptual categories are defined as points in a psychophysical space with continuous dimensions. When a listener hears a sound, this sound is evaluated on a number of dimensions (e.g., duration, frequency) and mapped onto a point corresponding to its values in multidimensional space. Sounds

originating from distinct categories are consistently mapped to the same points and repeated exposure to these categories leads to the formation of distinct "clouds" that listeners can start to associate with a phonetic category.

We assume that, in essence, auditory category learning is equivalent to recognizing the statistical patterns that are present in the signal (Pierrehumbert, 2003). For example, exposure to the stimulus structure in the upper left panel of Figure 4.1 should encourage listeners to categorize using only dimension 1 and ignore dimension 2, whereas exposure to the stimulus structure in the lower left panel should encourage listeners to categorize using only dimension 2 and ignore dimension 1. 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.

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 to start using more than one dimension in their categorizations (Ashby, Alfonso-Reese, Turken, & Waldron, 1998). Ashby et al. (1998) argue that there are two category learning systems, a verbal learning system and a procedural based 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 high frequency sounds in category A" only enter the verbal system after all unidimensional options have been tried. The other category learning system is a procedural or implicit learning system that does not have the

same preference for unidimensional rules. This system is also not as dependent on feedback as the verbal system for learning but learns much more slowly.

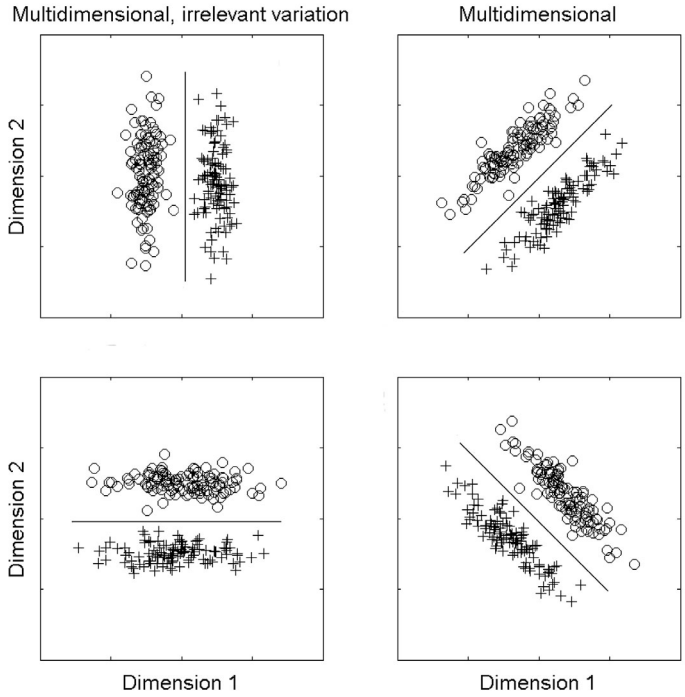

Figure 4.1. Four possible category structures in a two-dimensional perceptual space. Lines represent the optimal solution to the categorization problem.

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 category learning. Ashby, Queller, and Beretty (1999) showed the initial preference of listeners for unidimensional solutions (using one dimension and ignoring the other(s)) in unsupervised learning. Only when their subjects got trial-by-trial feedback they were able to learn a distinction based on more than one dimension. Homa and Cultice (1984) also showed the preference of subjects for relatively simple (easily verbalizable) distortions of dot-pattern stimuli

in unsupervised learning. When the distortions got too large, subjects were unable to classify the stimuli correctly without supervision. Furthermore, Love (2002) showed that learning performance with unidimensional categorization problems in unsupervised learning is far superior to their performance with more complex problems.

Considering the predictions of the Perceptual Assimilation Model, the native language of the listeners becomes an important issue in any perceptual learning experiment. The cleanest case for the study of second language acquisition would be the situation where assimilation does not happen but sounds are still recognized as speech in the native phonological space. This is the case when the sounds are located in a relatively empty area of phonetic space. We argue that this is the case for Spanish listeners and the Dutch high front vowels /ʏ/ (as in /fʏt/, "fut"; "energy"), /y/ (as in /fyt/, "fuut"; "grebe") and /ø/.(as in /føt/, "feut"; "freshman"). These vowels differ from each other primarily in the frequency of their first formant (formant frequency) and their duration. The sounds /ʏ/ and /y/ do not differ greatly in length, but /y/ has the lower first formant frequency, while the sounds /ø/ and /ʏ/ have similar spectra but /ø/ has a longer duration.

The Spanish language has a relatively small vowel inventory of five vowels: /i/, /e/, /a/, /o/, and /u/. These vowels differ in height, backness and roundedness (Hammond, 2001; Bradlow, 1995). These articulatory dimensions correlate with the first and second formants of the acoustic signal in a F1/F2 vowel space. 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 unrouded vowels (/e/ and /i/) have a high value for F2 (Bradlow, 1995).

Dutch has a large vowel inventory of sixteen simple vowels and three diphthongs (Booij, 1995). An important difference with Spanish for our purposes is the existence of a durational contrast between certain vowels in Dutch, for example, /ʏ/ is a short version of /ø/. Furthermore, while the Spanish vowels are all situated at the outside of the F1/F2 vowel space, Dutch also has some vowels situated in the center of this space, notably, /ʏ/ and /ø/. These vowels thus constitute an example of Best's Uncategorized case for a Spanish listener because they occupy an empty part of Spanish vowel space. They are too far removed from any native Spanish vowel category to be assimilated.

In Experiments 3 and 4, listeners of a language with a bigger vowel inventory, American English, categorize the same vowels as the Spanish listeners in Experiments 1 and 2. Although American English, like Dutch, has a large vowel inventory with fifteen vowels (Ladefoged, 1999), the area in vowel space that corresponds to the three Dutch vowels /ʏ/, /y/, and /ø/ is empty. All are unknown sounds in American English.

All experiments had a similar design with a pretest, a learning phase and a maintenance phase. The first panel of Figure 4.2 shows the distributional structure of the pretest. The stimuli are drawn from an equidistant grid with an equal range of variation in both stimulus dimensions. In the pretest, this grid is intended to neutrally scan the listener's initial categorization tendencies.

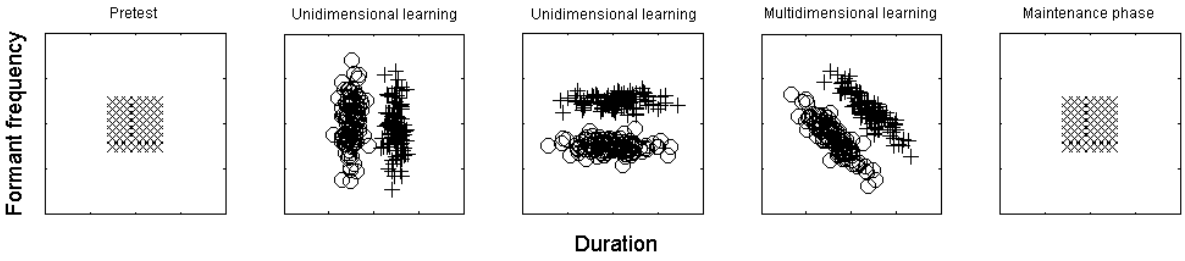

Figure 4.2. The basic experimental design of the experiments in this chapter: a pretest phase without distributional information, training phases with distributional information (either one or two relevant dimensions) and a maintenance phase that is identical to the pretest.

The second through fourth panel show the learning phases of the various experiments. The second and third panel depict category structures called "unidimensional learning" These consist of one relevant dimension of variation and one irrelevant dimension of variation. For optimal performance, listeners have to learn to use one dimension in their categorizations and learn to ignore the other dimension. In the second panel, listeners are trained to use duration as a relevant dimension, while in the third panel, listeners have to learn to use formant frequency in their categorization while simultaneously learning to ignore duration. This type of category structure is contrasted with that in panel four ("Multidimensional learning"), where both dimensions exhibit relevant variation. For optimal performance, listeners have to learn to use both dimensions in their 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 phase 1 and learning phase 2) to investigate possible changes in categorization behavior over time.

All experiments ended with a maintenance phase that had the same stimuli as the pretest and was again intended to scan the listeners' perceptual space in the

absence of distributional information. If listeners learned a new category structure in the learning phase and if they are able to transfer this learning to the maintenance phase, performance in the maintenance phase should resemble that of the learning phase, especially in comparison with performance in the pretest.

The experiments presented in this chapter investigate the role of the native phonology, the role of the distributional information provided to listeners and the role of supervision in phonetic category learning. Experiment 1 investigates supervised learning of Spanish listeners who are being trained either on the distinction between /ø/ (longer duration) and /ʏ/ (shorter duration), where duration is the relevant dimension of variation or on the distinction between /ʏ/ (higher F1) and /y/ (lower F1), where formant frequency is the relevant dimension of variation. Experiment 2 investigates unsupervised learning of the same distinction. To investigate the possible role of the native phonology, Experiment 3 examines the difference between Spanish and American English listeners learning the distinction between /ø/ (longer duration) and /ʏ/ (shorter duration) with the aid of trial-by-trial feedback. Experiment 4 trains American English listeners with the aid of trial-by-trial feedback on the distinction between /ø/ (longer duration and high F1) and /y/ (shorter duration and lower F1). To categorize the stimuli successfully, listeners will have to use both dimensions in their categorizations, something that has been shown to be difficult for listeners of various language groups (Flege & Hillenbrand, 1986).

Experiment 1

Method

Subjects[12]

Twenty (ten in each condition) Spanish exchange students from the Radboud University of Nijmegen participated in the experiment. None of them spoke another language besides English, but most of them were engaged in learning Dutch. Their proficiency in Dutch was extremely low. All listeners reported normal hearing. After the experiment they filled in a questionnaire about their listening experiences to assess whether the stimuli were recognized as vowels. All listeners qualified the stimuli as such.

Stimuli

The categories of both conditions had one relevant dimension of variation (see the second and third panel of Figure 4.2). In condition 1, the variation in duration was relevant, whereas formant frequency varied irrelevantly. The means of the two categories corresponded roughly 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). In condition 2, the duration was kept constant and formant frequency was systematically varied. The means of the two categories corresponded roughly 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

12 Laurence Bruggeman is kindly acknowledged for her assistance in recruiting and testing the Spanish listeners.

each other primarily in the frequency of their first formant (formant frequency) with ʏ being a higher (more frontal) version of/ʏ./ (Booij, 1995). All vowels occur frequently in Dutch and were synthesized using the PRAAT Speech synthesis program (Boersma, 2001).

Careful listening by native Dutch listeners 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 the dimensions (see the rightmost panel of Figure 4.2). 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 to the number of jnds of the frequency range.

Table 4.1 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.

Table 4.1.

Stimulus characteristics of the phonetic categories used in Experiment1, 2, and 3. The rows presenting the learning stimuli of Condition 1 (duration relevant) and Condition 2 (formant frequency relevant) list the mean stimulus duration and standard deviation in D and in ms and the mean value of the first formant and its associated standard deviation in ERB and in Hz. Any deviation of correlation coefficient ϱ from 0 is due to sampling. Both conditions have the same maintenance phase stimuli. The mean, minimal, and maximal values of both duration and formant frequency of the maintenance stimuli 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 to F5 used in the generation of all stimuli. Bandwidths were set at10 % of the frequency.

	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*		
	Means	σ	ϱ	Means	σ	ϱ
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 1.8ERB 120 Hz	-0.10
	Maintenance stimuli					
	Mean		Min	Max		Stepsize
Duration	51.1 D 131 ms		50.0 D 101 ms	52.2 D 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

Procedure

Listeners were seated in a soundproof booth in front of a computer screen and a two-button response box. The listeners' task was to assign each stimulus to group A or B, using the two-key button box.

The experiment again consisted of a pretest, two learning phases and a maintenance phase. Using a pretest allowed us to detect any preexisting categorization tendencies. The pretest 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 Figure 4.2). 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 saw fit at the end of the learning phase.

The learning consisted of 448 stimuli (2 categories times 2 repetitions times 112 stimuli per category) presented at a comfortable level through Sennheiser headphones (HD 270). The stimuli from the two categories were presented in a 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. Listeners had to assign the learning stimuli to category A or B with the two-key button box. Once participants had selected a category label on a trial and their categorization was correct, the monitor displayed (the Spanish equivalent of) "right" in green letters for 700 ms; when the categorization was incorrect, the monitor displayed (the Spanish equivalent of)

"wrong" in red letters for 700 ms immediately following the response. After the visual feedback disappeared, a 200 ms blank screen preceded the next stimulus.

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

Results and discussion

Signal detection analysis

As a first analysis, percent correct and d' were calculated for the learning phases of each condition (See Figures 4.3 and 4.4 and Table 4.2). Recall that in the pretest and maintenance phase a stimulus grid was used without feedback, so correct and incorrect categorization did not apply in these phases. Pretest and maintenance phases are analyzed in detail in a later section. Figure 4.3 and 4.4 suggest a learning effect, judging by the increase in performance from the first to the second learning phase.

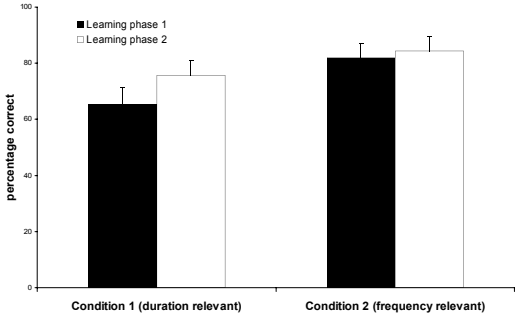

Figure 4.3. Percentage correct for the two learning phases of Conditions 1 and 2 of Experiment 1.

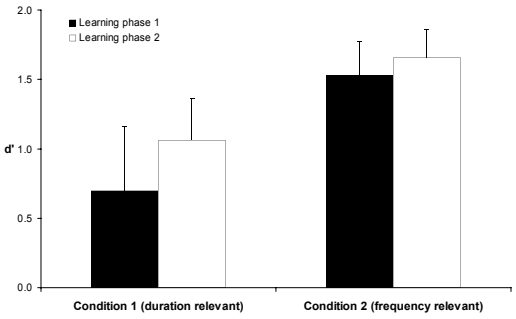

Figure 4.4. d' values of the two learning phases of Condition 1 (duration relevant) and 2 (frequency relevant) of Experiment 1

After having confirmed that percentage correct differed significantly from chance in all phases of the experiment (minimum t [9] = 2.92, $p < 0.05$), we tested the learning effect using an ANOVA with Part of the Experiment as within-subjects variable and Condition as between-subjects variable. This analysis showed the percentage correct to be significantly higher ($F [1,18] = 6.30$, $p < 0.05$) in the second learning phase. This effect did not interact with Condition, so the learning was equal in both conditions. Further, the analysis showed there to be a marginally significant advantage for Condition 2($F [1,18] = 3,066$, $p < 0.097$), where formant frequency was the relevant dimension, meaning that subjects tended to be better at categorizing stimuli when formant frequency was the relevant dimension than when duration was the relevant dimension.

In all phases of the experiment and for both conditions, d' differed significantly (minimum t [9] = 1.89, $p < 0.05$) from zero (the value associated with identical distributions of perceptual effects of two stimuli in signal detection theory, (Macmillan & Creelman, 1997). As with percentage correct, the main effect of Part of the experiment was significant for the d's ($F [1,18] = 7.58$, $p < 0.05$). The difference between Conditions was again marginally significant ($F [1,18] = 4.08$, $p < 0.06$). There was no significant interaction between Condition and Part of the experiment.

Table 4.2.
Signal detection analysis results for Experiment 1 (supervised learning with relevant variation in one dimension and irrelevant variation in the other dimension). The mean percentage correct and d' values and their associated standard deviations are displayed for both learning phases of Conditions 1 and 2.

	Learning phase 1				Learning phase 2			
	pc	σ	d'	σ	pc	σ	d'	σ
Condition 1 (duration relevant)	0.66	0.17	0.70	0.83	0.76	0.18	1.06	0.95
Condition 2 (formant frequency relevant)	0.82	0.17	1.53	0.77	0.84	0.16	1.66	0.65

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

Logistic regression

The binary choice design (every answer is either category A or category B) is very well suited by a logistic regression. A logistic analysis yields two β-weights (which can be significant or not) which indicate 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.

Figure 4.5 and Table 4.3 show mean β-weights, standard errors (Figure) and standard deviations (Table) of the dimensions duration and formant frequency for the pretest ("Pretest"), the first and second learning phase ("Learning phase 1" and "Learning phase 2") and the maintenance phase ("Maintenance phase").

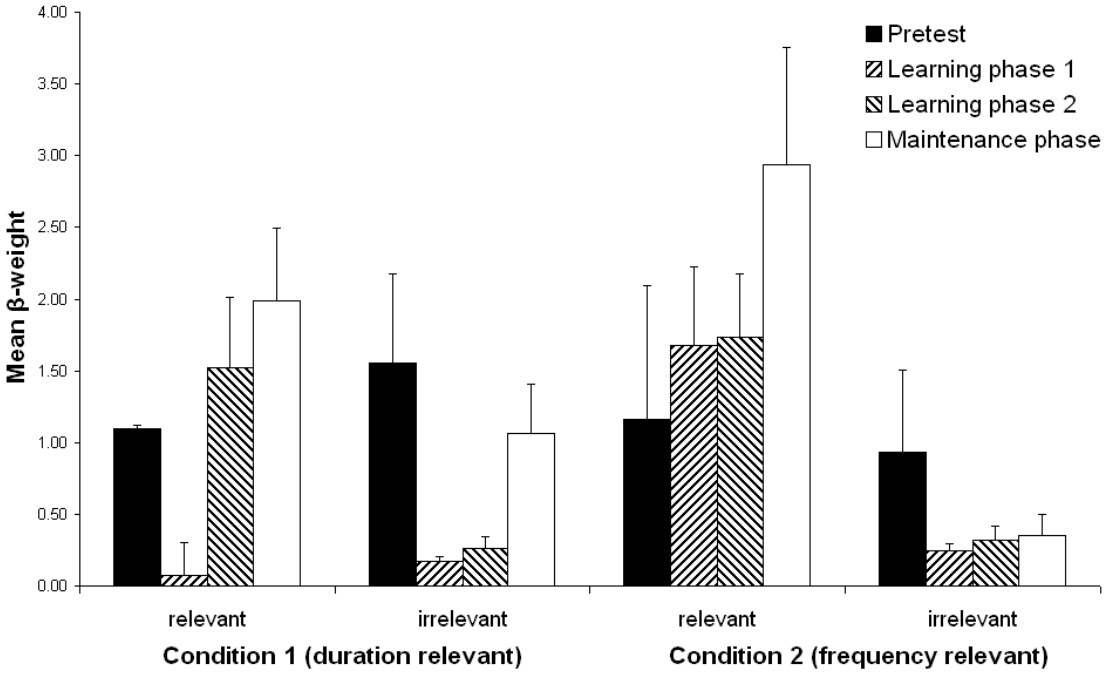

Figure 4.5. Mean β-weights and their respective standard errors of the relevant and irrelevant dimensions for Condition 1 (duration relevant) and Condition 2 (frequency relevant) for each Part of the experiment.

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 labeled "Uni" and "Multi" of Table 4.3 show how many subjects either used one or all dimensions significantly. These categories are mutually exclusive and subjects using neither dimension have been omitted.

Table 4.3.

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

	Pretest							
	Condition 1, duration relevant (N=10)				Condition 2, F1 relevant (N=10)			
	μ(β)	σ(β)	Uni	Multi	μ(β)	σ(β)	Uni	Multi
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							
	μ(β)	σ(β)	Uni	Multi	μ(β)	σ(β)	Uni	Multi
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							
	μ(β)	σ(β)	Uni	Multi	μ(β)	σ(β)	Uni	Multi
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							
	μ(β)	σ(β)	Uni	Multi	μ(β)	σ(β)	Uni	Multi
Relevant	1.99	1.61	4	3	2.94	2.58	5	2
Irrelevant	1.06	1.10	2		0.35	0.46	1	

Figure 4.5 and Table 4.3 both 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 dimensions were higher than those for the irrelevant dimensions. There were some differences between Conditions 1 and 2, possibly reflecting a preference for formant frequency as a relevant dimension. First, the β-weight for the relevant dimensional in Condition 1 was low in the first learning phase, reflecting listeners' reluctance to use

this dimension. Similarly, ignoring duration in the maintenance phase when it is irrelevant (Condition 2) appears to be easier than ignoring formant frequency in the maintenance phase when it is irrelevant (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 both the increase in mean β-weight as the experiment progressed ($F [3, 54] = 9.096$, $p < 0.05$) and in the overall preference for the relevant over the irrelevant dimension ($F [1,18] = 7.86$, $p < 0.05$). Performance in Condition 2 was not better than performance in Condition 1 ($F [1, 18] = 0.17$, *n.s.*). The initial preference of our listeners for formant frequency led to a significant interaction between Part of the experiment and Dimension ($F [1,54] = 7.45$, $p < 0.05$) with formant frequency always being the preferred dimension in the Pretest.

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. Independent of whether a relatively unfamiliar dimension (recall that duration does not play a significant role in the Spanish vowel system) is relevant or a very familiar one (formant frequency), our listeners were sensitive to the information provided to them and could maintain the distinction they learned in the maintenance phase. The Dutch listeners of Chapter 2 and 3, much more familiar with the dimension duration, preferred to use duration in the maintenance phases.

The trial-by-trial feedback provided in this experiment is not often available to the second (or first) language learner. Usually, when learning a second language, we have to rely on the same distributional information available to infants learning a

first language, with the possible inclusion of some lexically driven information (Eisner & McQueen, 2005) although this type of perceptual learning has only been shown to be able to fine-tune listeners' categories, not to create new ones. In Experiment 2, unsupervised learning of the same speech categories as in Experiment 1 is investigated.

Experiment 2

Method

Subjects

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

Stimuli

Both the learning and pretest/maintenance stimuli were identical to those used in Experiment 1 (see Table 4.1). Synthesized versions of the Dutch vowels from the words "fut" (/fʏt/), "feut" (/føt/) and "fuut" (/fyt/). In Condition 1, the relevant dimension of variation was duration and in Condition 2 the relevant dimension of variation was formant frequency.

Procedure

The procedure in the pretest and maintenance phase was identical to that in Experiment 1 (See the "unidimensional learning" panels of Figure 4.2). Contrary to the procedure of Experiment 1, no trial-by-trial feedback was provided in the learning phases. 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.

Results and discussion

Signal detection analysis

The signal detection measures percent correct and d' are presented in Table 4.4and in Figures 4.6 and 4.7 respectively.

Table 4.4.
Signal detection analysis results for Experiment 2 where Spanish listeners had to learn to categorize stimuli with relevant variation in one dimension and irrelevant variation in the other dimension without supervision. The mean percentage correct and d' values and their associated standard deviations are displayed for both learning phases of Condition 1 (duration relevant) and 2 (formant frequency relevant).

	Learning phase 1				Learning phase 2			
	pc	σ	d'	σ	pc	σ	d'	σ
Condition 1 (duration relevant)	0.61	0.16	0.71	0.89	0.66	0.19	0.56	0.69
Condition 2 (formant frequency relevant)	0.84	0.13	1.91	1.12	0.85	0.13	1.92	1.12

The figures as well as the table indicate a better performance in Condition 2 compared to Condition 1. The figures show little indication of the learning effect found in Experiment 1 in the difference between the learning phases.

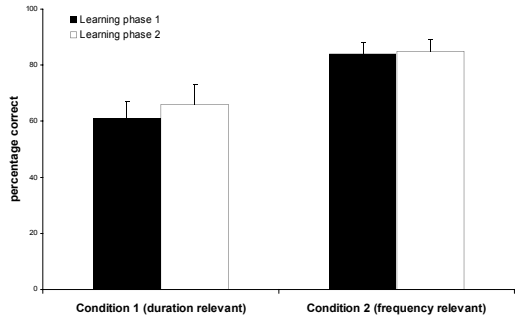

Figure 4.6. Percentage correct for the two learning phases of Conditions 1 and 2 of Experiment 2.

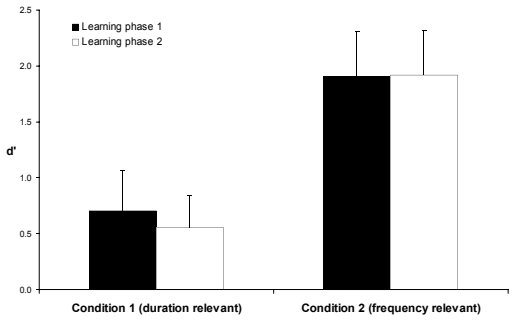

Figure 4.7. d' values of the two learning phases of Condition 1 (duration relevant) and 2 (frequency relevant) of Experiment 2.

Before statistically testing these observations we first tested whether the percent correct scores differed significantly from chance. The chance level for an experiment without trial-by-trial feedback is less obvious than in an experiment with supervision. In order to calculate percent correct, the listener's response must be labeled “right” or “wrong”, depending on whether he or she assigns a stimulus to the correct category. 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 based on his or her performance. 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, chance level is not simply at 50% correct. We calculated the expected value for chance level for 224 stimuli from a binomial distribution and the transformed percent correct, leading to a test value of 52.66%.

In Condition 1, when duration was the relevant dimension, percentage correct did not differ from chance in the first learning phase (t [5] = 1,65, n.s.) or in the

second (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 was also present in the main effect for Condition in the ANOVA (F [1,12] = 7,77, $p < 0.05$) with Part of the Experiment as independent within-subject variable. There was no significant effect of Part of the experiment (F [1,12] = 0.012, n.s.) in Condition 1.

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 phase 1 (t (7] = 4.83, $p < 0.05$) and phase 2 (t [7] = 4.84. $p < 0.05$) differed significantly from zero. The d's in Condition 2 were also 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, a significant effect of Condition (F [1,12] = 5.85, $p < 0.05$)) was found, in the absence of an effect of Part of the experiment or an interaction.

These analyses show that performance was good when formant frequency was the relevant dimension, but not when duration was relevant. These effects will be further explored in the logistic regression analyses.

Logistic regression

Table 4.5 and Figure 4.8 show the mean β-weights of Condition 1 (duration relevant and Condition 2 (formant frequency relevant) for all four phases of the experiment.

Table 4.5.

Logistic regression results of Experiment 2 where Spanish listeners were had to learn to categorize stimuli with relevant variation in one dimension and irrelevant variation in the other dimension without supervision.. 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.

	Pretest							
	Condition 1, duration relevant (N=6)				Condition 2, F1 relevant (N=8)			
	μ(β)	σ(β)	Uni	Multi	μ(β)	σ(β)	Uni	Multi
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							
	μ(β)	σ(β)	Uni	Multi	μ(β)	σ(β)	Uni	Multi
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							
	μ(β)	σ(β)	Uni	Multi	μ(β)	σ(β)	Uni	Multi
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							
	μ(β)	σ(β)	Uni	Multi	μ(β)	σ(β)	Uni	Multi
Relevant	1.38	1.96	2	0	3.20	2.30	6	0
Irrelevant	0.96	1.25	2		0.68	0.87	2	

Unsupervised learning of category structures with relevant variation in only one dimension appears to be difficult. With the variable dimension coded as "relevant" versus "irrelevant", there was no significant effect of dimension ($F [1,12] = 0.345$, *n.s.).* This means that participants did not show an overall preference for the relevant dimensions over the irrelevant one; 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, as significant interactions between Part of the experiment and Condition ($F [3,36] = 7,25$, $p < 0.05$) and Part of the Experiment and Dimension ($F [3,36] = 3,93$, $p < 0.05$) indicate. To further investigate this, separate analyses were conducted for each condition and each combination of pretest/maintenance phase and the two learning phases. This showed that the interactions were carried by the interaction between the Dimension and Part of the experiment (Pretest versus Maintenance phase) of Condition 2 ($F [1,7] = 7,928$, $p < 0.05$). Only when formant frequency was the relevant dimension, it was used more in the maintenance phase compared to the irrelevant dimension in the pretest.

Although the differences between supervised and unsupervised learning are considerable, an overall ANOVA with Supervision, Dimension, and Condition as between-subjects variable and Part of the Experiment as within-subjects variable, failed to show a significant main effect of Supervision ($F [1,33] = 0.27$, *n.s.*) nor was there any relevant interaction.

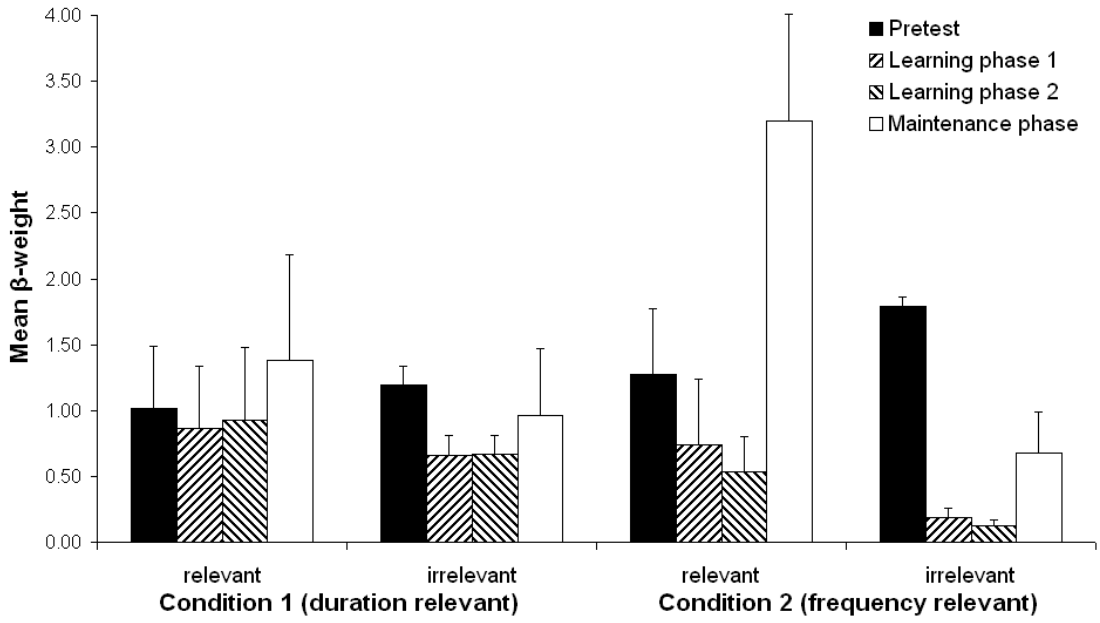

Figure 4.8. Mean β-weights of Experiment 2 for Condition 1 and Condition 2 for the relevant and irrelevant dimensions for each part of the experiment. In Condition 1, duration was the relevant dimension; in Condition 2, formant frequency was relevant. Vertical line segments indicate plus one standard error.

The results of both the supervised and unsupervised learning experiment indicate a preference of our Spanish listeners for the dimension of frequency of the first formant. Especially when considering the percentage correct levels, performance was better when formant frequency was the relevant dimension. We hypothesize this was because of the phonological structure of the language, where duration is not an important phonetic dimension in distinguishing vowels whereas formant frequency is (Hammond, 2001).

A study by Kawahara (2006) with Japanese and English listeners has shown that the duration of non-speech stimuli can be perceived differently by listeners with differing phonologies. Thus we next tested listeners whose phonology differed from that of the Spanish listeners in Experiments 1 and 2; we examined supervised

learning of the stimuli of Condition 2 by speakers of American English. While duration may not be a strict phonetic cue in American English, there is much more variation in the average duration of these vowels (Hillenbrand, Getty, Clark, & Wheeler, 1995), the distinction between tense and lax vowels in English is associated with (allophonic) duration differences with tense vowels being longer and lax vowels being shorter (Smiljanić & Bradlow, 2005), and vowel duration signals the difference between some voiced and voiceless consonants (Flege & Hillenbrand, 1986). If the performance of the American English listeners betters that of our Spanish listeners, this would be evidence of the importance of the native phonological system in learning new phonetic categories.

Experiment 3

Method

Subjects[13]

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

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

Stimuli

The stimuli were identical to those used in Condition 1 of Experiments 1 and 2. Thus, duration was the relevant dimension of variation for categorizing the stimuli, while formant frequency varied irrelevantly. See Table 4.1.

Procedure

The procedure was similar to that of Experiment 1. The experiment again consisted of a pretest, two learning phases and a maintenance phase. After the listeners had received instructions and signed consent, they were seated in a soundproof booth and pressed a button to start the experiment. 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, in the maintenance phase subjects had to try to maintain the rule 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 presented twice, resulting in 448 trials. The learning phase lasted for about 25 minutes, depending on the response speed of the subjects.

After the learning phase, listeners categorized the pretest stimuli again in the maintenance phase according to the rule they had discovered in the learning phase. Finally, all participants filled out a questionnaire asking them whether they

recognized the sounds as speech, whether they labeled the groups in any way and whether they spoke a language besides English.

Results and discussion

Signal detection analysis

Again, the percent correct and the d' were calculated for each condition and part of the learning phase. The upper part of Table 4.6 lists the values for d' and the percent correct for Experiment 3. See also Figure 4.9 and 4.10.

Table 4.6.
Signal detection analysis results for Experiment 3 and 4 (American English listeners). The mean percentage correct, d' values and their associated standard deviations are displayed for both learning phases.

	Learning phase 1				Learning phase 2			
	pc	σ	d'	σ	pc	σ	d'	σ
Experiment 3 (duration relevant)	0.84	0.12	1.55	0.67	0.88	0.02	1.95	0.89
Experiment 4 (Multidimensional)	0.64	0.09	0.53	0.37	0.65	0.10	0.60	0.41

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. An ANOVA with language (Spanish versus English) as between-subjects variable and Part of the experiment (learning phase 1 versus learning phase 2) as within-subjects variable was conducted for both the percent correct and the d'.

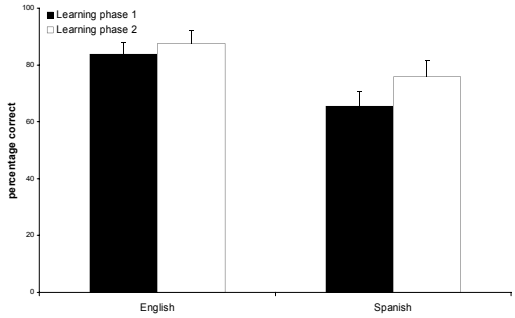

Figure 4.9. Percentage correct for the two learning phases of Experiment 3 (American English listeners) and Condition 1 of Experiment 1 (Spanish listeners). Learning was supervised and duration was the only relevant dimension of variation.

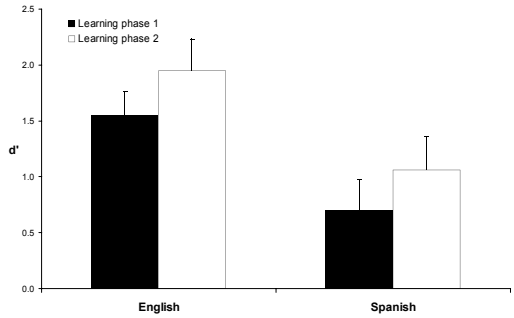

Figure 4.10. d' values for the two learning phases of Experiment 3 (American English listeners) and Condition 1 of Experiment 1 (Spanish listeners). Learning was supervised and duration was the only relevant dimension of variation.

The results show 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' ($F [1,18] = 5.45$, $p < 0.05$). Together with the absence of any significant interactions between Part of the experiment and Language, the significant main effect of Part of the experiment for both percent correct ($F [1,18] = 8.71$, $p < 0.05$) and d' ($F [1,18] = 33.57$, $p < 0.05$) show, however, that both language groups were able to learn to use the dimension duration.

Hence, there was a difference in performance measures between the two language groups. English listeners who are more familiar with distinguishing vowels based on duration due to their native phonology, performed better when they had to learn to categorize using duration.

Logistic regression

As in Experiments 1 and 2, a logistic regression analysis was performed, Figure 4.11 displays the results of this analysis for the pretest ("Pretest"), the first part of the learning phase ("Learning phase 1"), the second part of the learning phase ("Learning phase 2") and the maintenance phase ("Maintenance") of both the American English and Spanish language groups. Table 4.7 displays the mean β-weights, standard deviations as well as the number of subjects using a dimension significantly for each part of the experiment.

Figure 4.11 as well as the comparison between Table 4.7 and Table 4.5 clearly show the differences between the two languages. The mean β-weights for the relevant dimensions were higher for the American English listeners and the mean β-weights for the irrelevant dimension formant frequency were higher for the Spanish listeners. It seems that using the relevant dimension as well as suppressing an irrelevant one is more feasible when those dimensions are a part of the phonological structure of one's language. This interaction between relevance of the dimension and language ($F [1,18] = 4,55$, $p < 0.05$)[14] 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 shows the difficulty the Spanish listeners experience in suppressing the use of formant frequency when it is irrelevant.

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

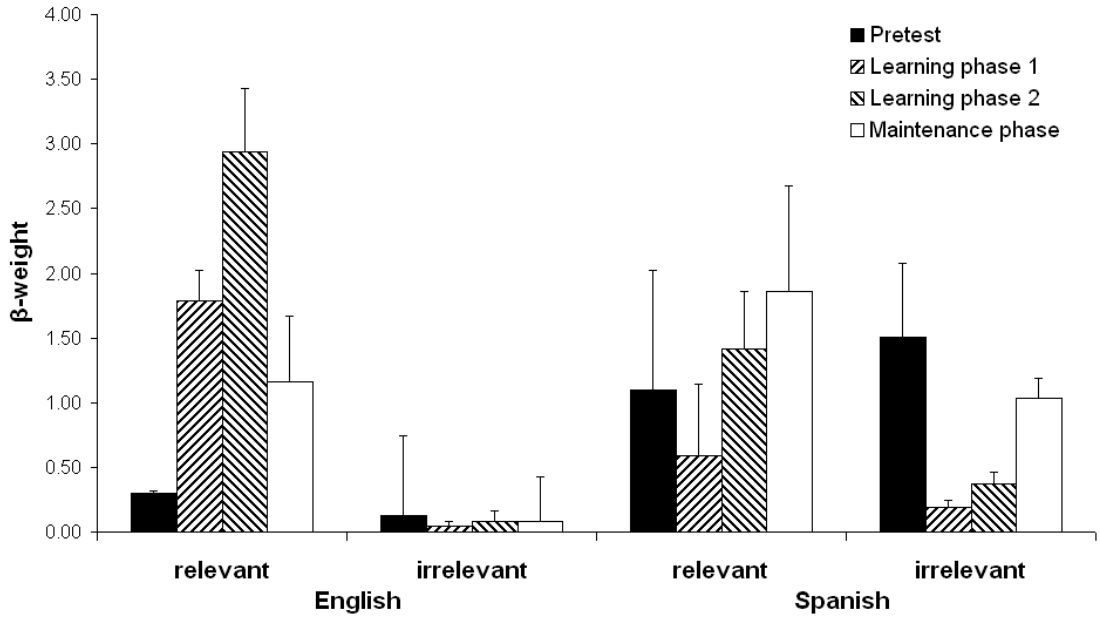

Figure 4.11. Mean β-weights for each part of the supervised learning experiment with duration as relevant dimension of variation for American English listeners (left bars) and Spanish listeners (right bars, taken from Condition 2 from Experiment 1).

Both dimensions showed a significant effect of Part of the experiment for the relevant dimension duration ($F [3,54] = 13.0$, $p < 0.05$) and for the irrelevant dimension frequency ($F [3,54] = 3,65$, $p < 0.05$). This main effect was modulated by Language in two significant interactions with Dimension for duration ($F [3,54] = 10.4$, $p < 0.05$) and for frequency ($F [3,54] = 3.10$, $p < 0.05$). This interaction again points to the differential preference of Spanish listeners for formant frequency. The lack of a significant Language effect for the relevant dimension 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 training phases are analyzed with an ANOVA with Language as between-subjects factor and Part of the experiment (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 ($F [,18] = 7.83$, $p < 0.05$), where Spanish has the higher β-weights.

Table 4.7.

Results of the logistic regression analysis of Experiment 3 where 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). 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.

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

The results of Experiment 3 and Condition 1 of Experiment 1 show the importance of the native language phonology in learning a new phonetic distinction. Both Spanish and American English listeners were able to learn the distinction based

on duration, but Spanish listeners experienced more difficulty with ignoring the irrelevant dimension formant frequency. American English listeners who were more familiar with the relevant dimension duration were better able to use this dimension and were also better able to ignore formant frequency.

In all experiments until now, learning was limited to situations where one dimension of variation was relevant and another dimension displayed irrelevant variation. This is in contrast to the situation with the phonetic inventory of most languages, where there is usually more than one relevant dimension of variation (Lisker, 1979). Furthermore, provided they are detectable, almost all aspects of the speech signal are considered relevant for phonetic categorization (Diehl & 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 when acquiring new phonetic categories (Flege & 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, both duration and formant frequency had to be used in distinguishing the categories.

Experiment 4

Method

Subjects

Eighteen undergraduate students from the University of Wisconsin, Madison participated in the experiment. All were native speakers of American English (and

so 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 questionnaire administered after the experiment were as in the previous experiments: all listeners judged the stimuli to be vowels or extremely like vowels.

Stimuli

Stimulus construction was identical to that in Experiment 1, except that the categories now had two relevant dimensions of variation (duration and formant frequency). See Table 4.8 for the stimulus characteristics of the learning phase. The pretest and maintenance stimuli were identical to those of Experiment 1, 2, and 3.

Table 4.8.
Stimulus properties of the multidimensional learning (Condition 2) stimuli of Experiment 4. 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 4.3

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	-0.95	113 ms	33.4 ms	-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 means of/ the two categories corresponded roughly 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 the sound (duration) were varied in creating the categories: /y/ is shorter and has a lower F1 than /ø/ (see the fourth panel of Figure 4.2).

Procedure

The procedure was identical to that used in Experiment 3: a pretest, two learning phases and a maintenance phase. In the pretest and the maintenance phase subjects were asked to categorize the stimuli into two groups. In the pretest listeners chose category labels as they wished, but in the maintenance phase subjects had to try to maintain the rule they had discovered in the learning phases. In the learning phase listeners received trial-by-trial feedback by lights above their response buttons. If a sound was not assigned correctly, the light belonging to the button that did signify the correct response would light up. In the learning phase, 112 stimuli times 2 repetitions times 2 categories were presented (448 stimuli). In the pretest and maintenance phase 49 stimuli were presented 4 times each (196 stimuli). The experiment lasted for about 40 minutes. Afterwards, all participants filled out a questionnaire asking them whether they recognized the sounds as speech, whether they labeled the groups in any way and whether they spoke a Germanic language besides English.

Results and discussion

Signal detection analysis

The second row of Table 4.6 shows the mean percentage correct and the mean d' for the first and second learning phase of Experiment 4. Figure 4.12 and 4.13 show the same data. The percentage correct and the d´ differed significantly from their respective chance levels (50% and 0) in all phases (min $t[17] = 6.10$, $p < 0.05$), but the difference between the first and second phase in the figures does not give a strong indication for a learning effect. Two ANOVA's with Part of the experiment as

within-subject variable and percentage correct or d' as dependent variables did not show a significant effect for either percentage correct (F [1,17] = 0.90, *n.s.*) or d' (F [1,17] = 0.30, *n.s.*).

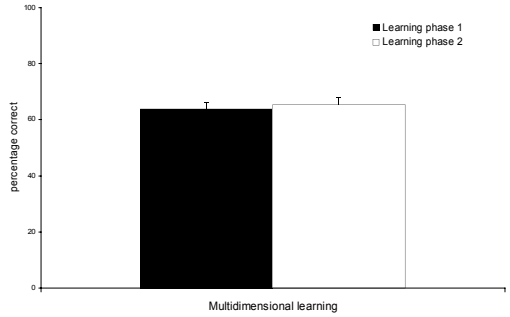

Figure 4.12. Percentage correct for the two learning phases of Experiment 4 (American English listeners). Learning was supervised and both duration and formant frequency were relevant dimensions of variation.

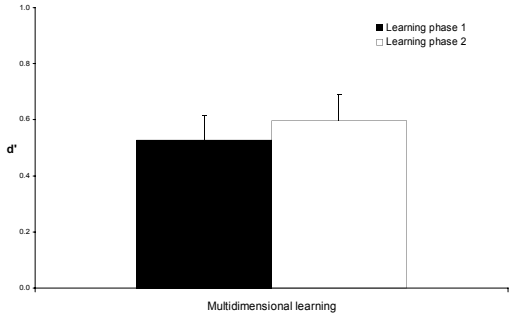

Figure 4.13 d' values for the two learning phases of Experiment 4 (American English listeners). Learning was supervised and both duration and formant frequency were relevant dimensions of variation.

The signal detection measures do not present any evidence of learning over time. Both measures, however, were significantly different from their chance levels, indicating that listeners were sensitive to the distributional information and the trial-by-trial feedback presented to them.

Logistic regression

The four panels of Figure 4.14 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 did not use any dimension significantly as circles.

Optimal performance corresponds to a point in the upper right hand corner of the Figure, with a ϕ of 45º (both dimensions are given equal weight) and far away from the origin (reflecting consistent behavior).

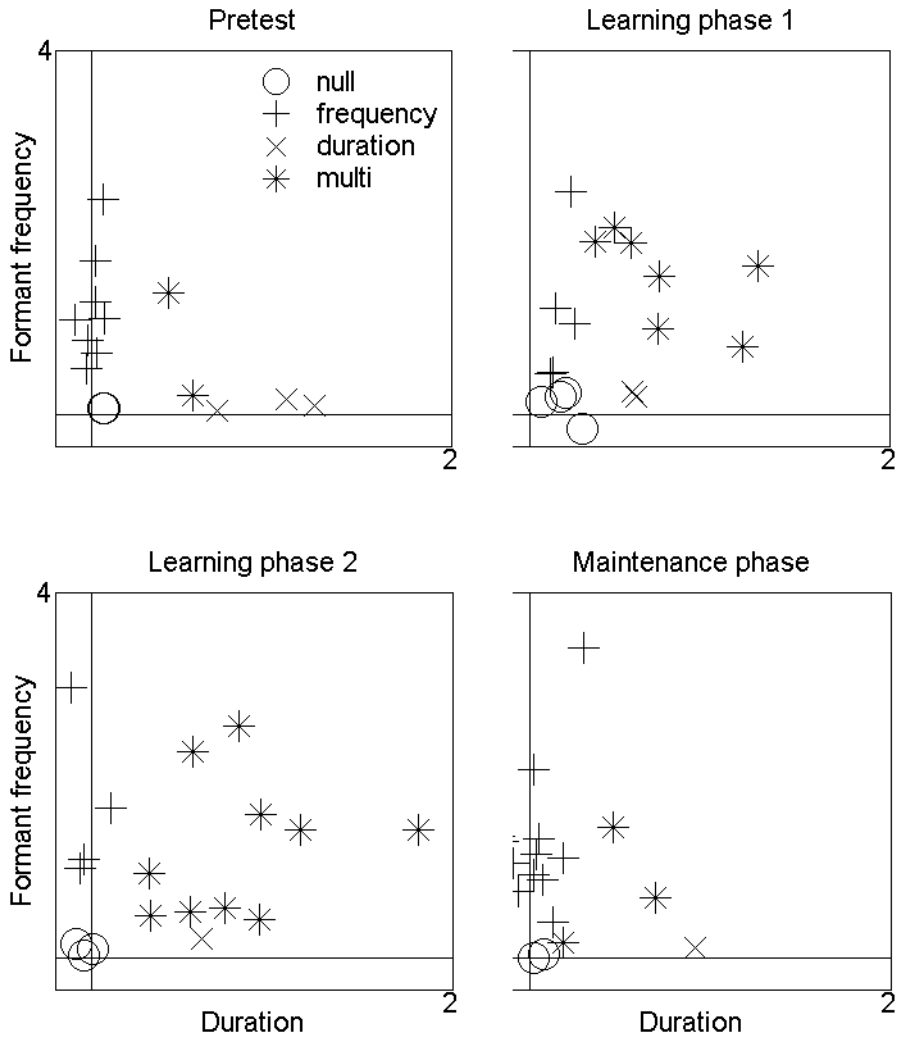

Figure 4.14. Scatterplots of individual β-weights for the two dimensions (duration or formant frequency) in Experiment 4 (two relevant dimensions of variation). Each of the four panels refers to a different part of the experiment.

The upper left panel of Figure 4.13 shows performance in the pretest. The majority of the listeners had a preference for using a unidimensional solution with

frequency (plus signs), which was also 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 who use both dimensions in their categorization increases (more asterisks), as does their consistency (asterisks further away from the origin). In the maintenance phase, when feedback was no longer given, much of this learning is 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, whereas others failed to use any dimensions 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 used 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 Figure 4.14), 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:

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

(5a) $\phi = \arctan(\beta_{freq}/\beta_{dur})$ if $\beta_{dur} \leq 0$

(5b) $\phi = \arctan(\beta_{freq}/\beta_{dur}) + \pi$ if $\beta_{dur} < 0$; $\phi - 2\pi$ if $\phi > \pi$

In our analysis, ϕ ranges between π and $-\pi$ radians. When ϕ equals $\frac{1}{2}\pi$, listeners purely use frequency, when ϕ equals 0, listeners use only duration, but when ϕ is close to $\frac{1}{4}\pi$ subjects are in between those two angles and use duration as well as frequency. As can be seen from Figure 4.14, most listeners fall in the upper right plane, somewhere between 0 and $\frac{1}{2}\pi$.

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 Figure 4.14, the listeners that categorized using both dimensions (indicated by the asterisks) are farther removed from the origin, while listeners that do not use any dimension significantly (the circles) are all very close to the origin. Table 4.9 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 $\frac{1}{2}\pi$ (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 $\frac{1}{2}\pi$ in all phases (min t [17] = 2.47, all $p < 0,05$). Although not all subjects categorize using a multidimensional rule, the subjects using formant frequency

balance those using duration, somewhat artificially resulting in an average multidimensional ϕ. Nevertheless, the number of listeners preferring the multidimensional solution over a unidimensional one increased during the learning phases, showing the capability of our listeners to profit from trial-by-trial feedback and distributional information.

Table 4.9.

Results of the logistic regression analysis of Experiment 4 where English listeners were trained with supervision on a category distinction where both dimensions were relevant. The angle ϕ, the consistency measure A as well as their respective standard deviations and the number of listeners significantly using one ("Uni") or both ("Multi") dimensions in their categorizations. Listeners using no dimension are not shown (N =- 18).

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

The consistency measure A was statistically evaluated in 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 using an ANOVA with Part of the experiment as within-subjects factor and Experiment (unidimensional versus multidimensional) as between-subjects factors. Performance in the unidimensional learning experiment was consistently better for both percentage correct ($F[1,26] = 24.67$, $p < 0.05$) and d' ($F[1,26] = 31.14$, $p < 0.05$).

Experiment 4 showed that listeners were sensitive to the distributional information and trial-by-trial feedback provided to them in this multidimensional category learning task. Compared to Experiment 3, however, performance in Experiment 4 was considerably worse. Learning a category distinction with more than one relevant dimension was considerably 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 considerable, but is probably insignificant compared to the exposure received by infants or 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 striking loss of this ability in the maintenance phase is similar to the loss of learned categorization skills observed in almost all the speech and non-speech learning experiments presented in this thesis. Listeners almost invariably prefer unidimensional solutions in category learning (Ashby, Queller & Berretty, 1999).

Although the learning phases of Experiment 4 showed that this preference can be modified, listeners reverted to a unidimensional categorization strategy in the absence of distributional information and trial-by-trial feedback.

General discussion

The experiments presented in this chapter investigated the processes involved in learning the sounds of a second language: the role of the phonological structure of the native language, the role of the distributional properties of the category distinction and the role of supervision in acquiring phonetic categories. The stimuli displayed tightly controlled variation in dimensions shown to be important in speech perception, duration and formant frequency. Depending on condition, this variation was either relevant or irrelevant to the category distinction.

Experiment 1 trained Spanish listeners to categorize non-native speech sounds with the aid of trial-by-trial feedback (supervision). The sounds listeners had to categorize varied on two dimensions, duration and formant frequency. Depending on condition, one of the dimensions was relevant whereas the variation in the other dimension did nothing to signal category membership. The results showed that listeners could learn to attend to the relevant dimension while suppressing the irrelevant one. The degree of success of this learning as well as its robustness in the maintenance phase depended heavily on which dimension was the relevant one. Learning and maintaining a distinction based on formant frequency was easier for our Spanish listeners than learning and maintaining a distinction based on duration.

Experiment 2 trained Spanish listeners to categorize the stimuli from Experiment 1 but now without trial-by-trial feedback. Listeners now had only one source of

information at their disposal: the distributional properties of the stimuli. With only this information available to them, performance levels decreased considerably and depended even more on whether duration or formant frequency was the relevant dimension. The results showed that without trial-by-trial feedback, listeners 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 best known to them, in this case formant frequency. Nevertheless, performance on several measures did differ from chance, showing the sensitivity of listeners to the distributional information.

Experiment 3 further tested the influence of the native phonology on categorization performance. American English listeners were presented with the same stimuli as in Condition 2 of Experiments 1 and 2 (duration relevant) and given trial-by-trial feedback in the learning phase. American English listeners are more acquainted with duration in their native phonology than Spanish listeners and have outperformed, we argue, the Spanish listeners on the signal detection measures as well as on the β-weights from the logistic regression.

Finally, Experiment 4 had listeners learn a category distinction with two relevant dimensions. Although the American English listeners were acquainted with both dimensions and received trial-by-trial feedback on their categorizations, performance was considerably impaired compared to supervised learning of a unidimensional distinction. Nevertheless, as with the impaired performance in the unsupervised learning of Experiment 2, listeners were certainly sensitive to the distributional information provided to them and the majority of listeners learned to use both dimensions in the categorization.

Taken together, the results show supervised learning to be superior to unsupervised learning, even when there is only one dimension of variation (Experiment 1 versus Experiment 2). The results also show that learning to categorize a category structure with one relevant dimension of variation and one irrelevant dimension of variation is more feasible than learning to categorize a category structure with two relevant dimensions of variation (Experiments1 and 3 versus Experiment 4). Nevertheless, even with as little as a few hundred presentations, listeners were shown to be 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. Finally, the results show the importance of the native phonology in learning a new category distinction. Although the distinction between /ʏ/ and /ø/ was new to both the American English and Spanish listeners, the American English group were much better at learning to categorize these two vowels. We argue that this is because the American English group is more acquainted with duration as a phonological dimension from their native phonology. The influence of the native phonology was also apparent in the preference of the Spanish listeners for formant frequency, especially in the pretest and maintenance phases when no distributional cues were present.

Comparing the results obtained in Chapters 2 and 3 with the previous results, we note a remarkable resemblance between the learning of auditory non-speech categories and the learning of phonetic categories. In both cases, supervised learning is superior to unsupervised learning and performance on category structures with one dimension of variation is significantly better than performance on category structures that require integration of two dimensions. Furthermore, listeners often

revert to their dimension of preference in the maintenance phase when distributional information is no longer present. This dimension of preference is shown to be dependent on the native language in the acquisition of phonetic categories. With non-speech auditory categories, the role of the native phonology cannot be determined because all listeners had the same native language. However, the preference for duration in the maintenance phases of Chapters 2 and 3 is also consistent with an important role of this dimension in Dutch phonology.

Chapter 5

Summary and conclusions

Summary

Acquiring phonetic categories for speech perception is more easily performed by infants than understood by adult researchers. This dissertation aimed at providing a better understanding of the processes involved in learning the categories of a first and a second language. The learning problem was operationalized as a two-category distinction involving one or two relevant acoustic dimensions. In particular, the role of supervision in the learning process, the role of the distributional information in the input and the role of the native phonology were investigated.

Non-speech category learning

The experiments presented in Chapter 2 investigated the supervised learning (defined as learning with trial-by-trial feedback) of non-speech categories. In Experiment 1 listeners were trained to categorize stimuli with one relevant dimension of variation and one irrelevant dimension of variation. The dimensions of variation in these experiments were always duration of the stimuli or the peak formant frequency of the stimulus. The results showed that these category structures were easy to learn with trial-by-trial feedback. However, maintaining the learned

distinction in a maintenance phase where this feedback was no longer present and the stimuli did not contain any distributional information anymore was considerably more difficult, especially when the relevant dimension was formant frequency and the irrelevant dimension was duration.

Supervised learning of a category structure where both duration and formant frequency were simultaneously relevant was investigated in Experiment 2 of Chapter 2. Learning such a truly multidimensional category distinction proved much more difficult, even with constant trial-by-trial feedback. Eventually, most, but not all, listeners mastered the distinction in the learning phase. This learning was far from robust, however, as in the maintenance phase listeners reverted to the use of only one dimension. As in Experiment 1, our listeners preferred duration over formant frequency in their unidimensional solutions of the maintenance phase.

Experiment 3 of Chapter 2 investigated whether the lack of trial-by-trial feedback or the absence of distributional information in the stimuli was responsible for participants' inability to maintain the category distinction in the maintenance phase. The learning phases of Experiment 3 were identical to those of Experiment 2, but the stimuli in the maintenance phase now were the same stimuli as in the learning phase. This time, listeners were able to maintain the multidimensional categorization strategy they had learned. Listeners were apparently very sensitive to the absence or presence of distributional information in the maintenance phase and adjusted their categorizations to suit.

The experiments presented in Chapter 3 investigated *unsupervised* learning (e.g., without trial-by-trial feedback) of the same category structures as in Chapter 2, again using non-speech stimuli. In Experiment 1 of this chapter, the categories exhibited relevant variation in one dimension and irrelevant variation in the other, as had been

the case in Experiment 1 in Chapter 2. The range of variation was equal for each dimension, so the only source of information for the listeners was the category structure of the stimuli. Irrespective of whether formant frequency or duration was the relevant dimension, listeners were able to determine the relevant dimension based on the distributional properties of the stimuli alone and use it in their categorizations.

In the learning phase of the experiment, performance was slightly better when formant frequency was the relevant dimension. In the maintenance phase, however, maintaining formant frequency as the relevant dimension and simultaneously ignoring duration was more difficult than vice versa. This preference for duration in the maintenance phase is similar to that found in the maintenance phase of the supervised learning experiment with the same stimuli in Chapter 2.

Experiment 2 of Chapter 3 investigated unsupervised learning of a category structure where both duration and formant frequency were relevant. Although the results were very variable, a significant proportion of listeners was sensitive to the distributional properties of the stimuli. However, benefiting from this multidimensional distributional information without trial-by-trial feedback proved much more difficult than using one dimension and ignoring another.

As was argued, the results of the unsupervised learning experiments from Chapter 3 do not differ qualitatively from the results of the supervised learning experiments from Chapter 2. In both learning situations, learning to categorize a category structure based on one relevant dimension of variation while ignoring the other was much more feasible than learning to use two dimensions simultaneously. Although comparatively difficult, learning a multidimensional category structure was possible, however, both with and without the aid of trial-by-trial feedback.

Supervised learning was quantitatively different from unsupervised learning; overall performance was better when learning was supervised.

Speech category learning

The experiments presented in Chapter 4 combined the supervised and unsupervised learning paradigms of Chapters 2 and 3. In contrast to the experiments from Chapters 2 and 3, synthesized Dutch vowels were used as stimuli instead of the non-speech sounds used in those experiments. These vowels, /ø/, /y/ and /Y/, can be distinguished from each other by using the same dimensions as were manipulated to create the non-speech sounds used in Chapters 2 and 3, namely duration and formant frequency. The listeners were speakers of Spanish (Experiments 1 and 2) and American English (Experiments 3 and 4), languages that do not use these vowels.

Experiment 1 showed that, with supervision in the form of trial-by-trial feedback, Spanish listeners were able to learn to categorize speech categories with one relevant dimension of variation and one irrelevant dimension of variation. Again, they could use duration as well as formant frequency as the relevant and irrelevant dimension, but with these listeners there was a preference for formant frequency. This preference was especially noticeable in the maintenance phases of the two conditions (duration relevant and formant frequency relevant). We speculate that this preference is due to the phonological properties of Spanish, where formant frequency is an important cue to vowel categorization whereas duration is not a relevant cue to category membership for vowels.

The preference for formant frequency was even clearer in Experiment 2, where learning was unsupervised and listeners had to rely solely on the distributional characteristics of the stimuli. Without supervision, Spanish listeners were not able to learn to categorize the two vowels based on one relevant and one irrelevant dimension. Their preference for formant frequency clearly showed in the maintenance phase of Condition 2 where formant frequency was the relevant dimension. Listeners used this dimension to great extent in their categorization, especially when compared to the use of duration in the maintenance phase of Condition 1 where duration was the relevant dimension but was hardly used in listeners' categorizations.

If the native phonology is responsible for the difficulty our Spanish listeners experienced in using duration in their categorizations, then speakers of another language that is more acquainted with duration in its vowel system should experience less difficulty in using duration in their categorizations. Experiment 3 tested this hypothesis by presenting American English listeners with the stimuli and paradigm of Condition 1 of Experiment 1 (supervised learning of a category structure with duration as the relevant dimension and formant frequency as the irrelevant dimension). Comparing the performance of the American English listeners with that of the Spanish listeners clearly showed an advantage for the English language group, thus supporting the hypothesis.

Experiment 4 of Chapter 4 investigated supervised learning by American English listeners of a category structure where both duration and formant frequency were relevant dimensions. The results showed that learning a multidimensional distinction is difficult for these listeners, even with the aid of supervision. Nevertheless, these listeners were shown to be sensitive to the two sources of

information (trial-by-trial feedback and the distributional properties of the signal) as more than half of them learned to use both dimensions by the end of the last learning phase. In the maintenance phase, just as in the multidimensional and unsupervised learning experiments of Chapters 2 and 3, listeners lost their ability to use both dimensions and reverted to a unidimensional solution, mostly preferring frequency as the dimension by which they categorized.

Conclusions

Taken together, the experiments presented in chapters 2 through 4 reveal several interesting and intriguing properties of auditory and phonetic category learning. Important conclusions can be drawn about the role of supervision and distributional properties in category learning, the similarities and differences between auditory and phonetic learning, the connection of the auditory category learning results with visual category learning results, the importance of sensitivity to distributional information in the category learning process and the differences between infant and adult auditory category learning. They will be discussed in the following sections.

Supervision and sensitivity to distributional information

First, success in acquiring auditory and phonetic categories depends both on the distributional properties of the categories and on the presence of absence of supervision. When there is only one relevant dimension of variation and the other dimension varies irrelevantly, listeners are well able to learn non-speech auditory categories. Whether learning of non-native speech sounds is possible, depends on

the presence or absence of trial-by-trial feedback. With feedback, our Spanish listeners are certainly able to learn a unidimensional distinction. Without feedback, learning a non-native unidimensional distinction is very difficult. For non-speech sounds, the presence or absence of feedback is relevant only for the degree of success. With trial-by-trial feedback, performance much better compared with unsupervised performance. However, even without trial-by-trial feedback, listeners certainly learn how to categorize the stimuli correctly.

Learning a multidimensional category distinction is much more difficult than learning a unidimensional one. Even with the aid of trial-by-trial feedback, listeners have a hard time mastering a distinction based on two dimensions. This holds for auditory category learning as well as for phonetic category learning. After being exposed to several hundred stimuli, listeners do show evidence of sensitivity to the relevance of both dimensions, but only sparsely so. When only one of these dimensions is relevant, they do not nearly experience as much difficulty. Without supervision, acquiring such a multidimensional category distinction is even more difficult.

This difficulty our listeners experience in learning a multidimensional speech or non-speech category distinction is surprising given the abundance of phonetic category distinctions that are based on more than one dimension. Perceptual and acoustic studies show that in order to reliably categorize vowel multiple dimensions are necessary (Hillenbrand, Getty, Clark, & Wheeler, 1995). There also is, however, a high level of redundancy in the signal listeners can use. Our multidimensional category structure lacked this redundancy. This difference between most speech category distinctions and our multidimensional category distinction, while necessary for our experiments, might have artificially made the learning of the

multidimensional category distinction more difficult. In this respect, the performance of our listeners is all the more impressive.

The importance of the distributional properties of the stimuli is also evident in the performance in the maintenance phases when there is no distributional information in the stimuli anymore. Although listeners were sometimes able to maintain the categorization strategy they learned, they also showed sensitivity to the absence of distributional information by starting to reuse the dimension that was irrelevant in the learning phases. The equidistantly spaced grid in the maintenance phases is designed to neutrally scan the listener's perceptual space and assigns equal weight to each dimension. Listeners apparently notice the equality and start using the irrelevant dimension again, especially when this is their preferred dimension.

The use of a maintenance phase without distributional information or trial-by-trial feedback is not common in visual category learning research but standard practice in phonetics and phonology. The difference in performance between the learning and maintenance phase has implications for speech research that uses similar equidistant continua to investigate the learning of phonetic categories (Repp & Libermann, 1987). The differential performance of our listeners in the learning phase and the maintenance phase is intriguing and shows the sensitivity of listeners to distributional information. What part of the distributional information is important is an empirical question. One suggestion is that listeners need the extreme stimuli that are present in the training distributions but not in the maintenance phase to keep the dimensions in mind and well calibrated.

This rapid adaptation to change in the input is reminiscent of the results found by Eisner & McQueen (2005). There, adults listeners were found to be extremely sensitive to changes in the pronunciation of the fricatives /s/ and /f/ and adjusted

their category judgments to suit. In contrast to our listeners, the listeners of Eisner and McQueen (2005) did have a lexical incentive to change their category judgments: perceiving a neutral fricative in one way (either /s/ or /f/) would result in a real word whereas perceiving the neutral fricative the other way would receive in perceiving a non-word. Our listeners stopped using their previously learned categorizations in the maintenance phase based only on the distributional properties of the stimuli.

Auditory and phonetic categories

A second important issue addressed by our experiments concerns similarities and differences between learning speech versus non-speech categories. Comparing the results from Chapters 2 and 3 with those of Chapter 4, the similarities are most striking. For both auditory and phonetic category learning a distinction based on one relevant dimension is easier to acquire than a distinction requiring the integration of two dimensions. Furthermore, the performance of our listeners in the maintenance phases without trial-by-trial feedback or distributional information was very similar for both speech and non-speech experiments: in both cases performance is usually not very robust and listeners tend to revert to a unidimensional solution involving a dimension of choice.

Auditory and phonetic category learning only differed considerably when there was no trial-by-trial feedback available in the learning phase. Then listeners were more sensitive to the distributional properties of the stimuli when they had to learn to categorize non-speech sounds. When Spanish listeners had to learn to categorize non-native speech sounds with one relevant dimension of variation and one irrelevant dimension of variation without the aid of supervision, they were unable to

determine which dimension was the relevant one, irrespective of whether this was duration or formant frequency. In the maintenance phases, formant frequency was shown to be the preferred dimension in their categorizations.

The similarities between auditory non-speech and phonetic category learning indicate that the complexity of our non-speech sounds is comparable to the complexity of speech sounds and as a result the non-speech sounds were, to a certain extent, also analyzed as such. This is an encouraging result, showing the possibility of conducting experiments relevant to the acquisition of speech categories with non-speech sounds in adults and thus avoiding all the possible interactions with the already present phonology of the native language (see also Mirman, Holt & McClelland, 2003).

Our results showed that in learning phonetic categories the native phonology is much more difficult to ignore than in learning auditory categories. This difference can be explained in terms of the Perceptual Assimilation Model. Although the listeners did consider the non-speech sounds to be speech-like in their complexity, they did not assimilate them into their native phonology. This situation is similar to the perceptions of the Zulu clicks by English listeners (Best, McRoberts and Sithole, 1988). Hence, the Dutch listeners categorizing non-speech stimuli were not as hampered by their preexisting phonological tendencies as the Spanish listeners were when did listened to non-native Dutch vowels.

Visual and auditory category learning

A third important aspect is the connection made between the literature on visual category learning (Ashby & Maddox, 1993; Nosofsky, 1990) and auditory category

learning research. Especially our method of stimulus construction and learning phases drew heavily from methods used in visual category learning. Compared to the difficult and complex category structures which subjects in visual category learning experiments are able to learn (Ashby & Gott, 1988; Ashby & Maddox, 1993), auditory category learning appears to be more difficult than the learning of visual categories.

However, the results also show similarities between auditory and visual category learning (Ashby, Alfonso-Reese, Turken & Waldron, 1998). The preference for a unidimensional solution is reminiscent of results from visual category learning (Ashby, Queller & Beretty, 1999) and the preference of listeners for a *particular* dimension in their unidimensional solutions is something also found in auditory category learning studies by Holt and Lotto (2005).

The initial preference for unidimensional solutions is abundant in visual category learning research. The category learning models COVIS (Ashby, Alfonso-Reese, Turken and Waldron, 1998) and SUSTAIN (Love, Medin and Gureckis, 2004) both incorporate this finding. In COVIS, the explicit (verbal) system is dominant over the implicit system (that is better able to learn multidimensional category distinctions). The explicit system is applied first and creates a verbal rule to describe the category distinction and verbalizing a unidimensional rule is easier than verbalizing a rule involving multiple dimensions. For example, the rule "Assign a sound to category A when its duration is less than 85 milliseconds" is less of a computational burden than the rule "Assign a sound to category A when its value for duration combined with its value for frequency does not exceed a criterion value". SUSTAIN also initially assumes a simple unidimensional category structure. Only when simple solutions are proven to be inadequate or when it is confronted with a surprising

event, an additional category is created and the category structure becomes more complex.

Considering our efforts to equalize the non-speech sounds by scaling their variability in empirically determined just noticeable differences, the preference of listeners for (in the non-speech case) duration is surprising. Just noticeable differences obtained in a same/different discrimination task apparently do not straightforwardly carry over to a category learning experiment. We explained the differential use of formant frequency and duration in terms of Stevens' and Galanter's (1957) *prothetic* dimensions (dimensions like duration where an increase in value means adding more of the same) and *metathetic* dimensions (dimensions like formant frequency where an increase in value does not necessarily mean more of the same). According to Smits, Sereno and Jongman (2006), the encoding of metathetic categories is noisier, hence the difficulty in maintaining the (weaker) representation of these dimensions in the maintenance phase.

For the speech stimuli, we argued that the differential preference for a dimension was based on the native phonology, something that was confirmed by the results of Experiment 3 from Chapter 4. Although the stimuli used in the experiments in Chapters 2 and 3 were non-speech, the native phonology could still play an important role. In Dutch, duration is a very important cue for vowel categorization and thus could have been the dimension of choice for our Dutch listeners.

Infant and adult learning of auditory categories

Fourth and finally, the difficulty our listeners experienced in learning without the aid of supervision and the lack of robust transfer to the maintenance phase makes

infant learning of phonetic categories (which is necessarily unsupervised) all the more impressive. Somehow, infants do succeed in robustly learning the sounds of their native language without the aid of supervision. Of course, the difference in amount of exposure might be an important factor - infants receive much more exposure than the 448 learning stimuli our listeners received. The Maye, Werken & Gerken (2002) study however, has shown that infants are capable of learning a new (unidimensional) phonetic distinction with only a few hundred exposures. Infants might have category learning skills that are not available anymore to adults.

In terms of the COVIS model (Ashby, Alfonso-Reese, Turken and Waldron, 1998), infants also do not have a verbal learning system that hinders adults in acquiring truly multidimensional auditory category distinction by trying to solve the category learning problem by searching for a unidimensional solution. The absence of a verbal category learning system is in this case beneficial to the infant because their implicit learning system can immediately start learning the multidimensional category boundary.

Related to the different situation of our adult listeners are the findings by Eisner and McQueen (2005) that showed that adults are able to shift their perception of speech sounds to suit the idiosyncratic or regional peculiarities of a given speaker. Adults listeners are perfectly able to slightly alter their phonetic category boundaries and maintain this information in memory (Eisner & McQueen, 2006), but our experiments showed that creating a totally new phonetic category in their already existing phonological space is considerably more difficult.

Why is this? In our daily life we constantly have to adapt to new speakers with different speaking habits and accents. Being able to shift your category boundaries is useful to adapt to these new speakers. Shifting a category boundary and trying to fit

an entire new category into the phonological space that is already divided into the phonetic categories of the native language, however, is a different understating entirely. A new category boundary that divides a native phonetic category into two non-native categories would conflict with categorical perception of the native language. It is thus arguable that the reason that adults find it so hard to learn new phonetic categories is that new categories in our native phonological space hinder the perception of our native language.

References

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

Agresti, A. (1990). *Categorical Data Analysis.* New York: John Wiley & Sons, Inc.

Ahn, W.K. & Medin, D.L. (1992). A two-stage model of category construction. *Cognitive Science*, *16*, 81-121.

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

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

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

Ashby, F.G. & Gott, R. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Human Perception & Performance*, *14*, 33-53

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

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

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

Ashby, F.G. & Maddox, W.T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149-178

Ashby, F.G. & Perrin, N.A. (1988). Towards unified theory of similarity and recognition. *Psychological Review*, 95, 124-150.

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

Ashby, F.G. & Townsend, J.T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*, 154-179.

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

Aslin, R.N., Jusczyk, P.W., & Pisoni, D.B. (1998). Speech and auditory processing during infancy: Constraints on and precursors to language. In D. Kuhn & R. Siegler (Eds.), *Handbook of Child Psychology*, Fifth edition. Volume 2: Cognition, Perception and Language (W. Damon, series editor) (pp 147-198). New York: Wiley.

Aslin, R.N., Pisoni, D.B., & Jusczyk, P.W. (1983). Auditory development and speech perception in early infancy. In M. Haith & J. Campos (Eds.), *Handbook of Child Psychology, Infancy and Developmental Psychobiology* (Vol. 2, pp. 573-687). New York: Wiley.

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

Best, C.T. (1993). Emergence of language specific constraints in perception of non-native speech: a window on early development. In B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. Mac-Neilage & J. Morton (Eds.) *Developmental Neurocognition: Speech and Face Processing in the First Year of Life* (pp 289-304).. Norwell, MA: Kluwer Acedemic Press.

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

Best, C.T., McRoberts, G.W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *Journal of the Acoustical Society of America*, *109*, 775-794.

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

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

Best, C.T. & Tyler, M.D. (2006). Nonative and second language speech perception: Commonalities and complementaries. In M.J. Munro & O.-S. Bohn (Eds) *Second language speech learning: the role of language experience in speech perception and production*. Amsterdam: John Benjamins.

Blumstein, S.E. & Stevens, K.N.(1979). Acoustic invariance in speech production: evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical. Society of America*, *66*, 1001-1017.

Blumstein, S.E. & Stevens, K.N.(1981).The search for invariant acoustic correlates of phonetic features. In P.D. Eimas and J.L. Miller, *Perspectives on the Study of Speech* (pp. 1-38). Hillsdale, NJ: Erlbaum.

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

Bradlow, A.R. (1995). A comparative study of English and Spanish vowels. *Journal of the Acoustical. Society of America*, *97*, 1916-1924.

Broersma, M. (2002). Comprehension of non-native speech: Inaccurate phoneme processing and activation of lexical competitors. *Proceedings of the 7th International Conference on Spoken Language Processing* (pp 261-264). Denver, Colorado.

Broersma, M. (2005). Perception of familiar contrasts in unfamiliar positions. *Journal of the Acoustical. Society of America*, *117*, 3890-3901.

Burnham, D.K., Earnshaw, L.J., & Clark, J. (1991). Development of categorical identification of native and non-native bilabial stops: Infants, children and adults. *Journal of Child Language, 18*, 231-260.

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

Cutler, A. & Broersma, M. (2005). Phonetic precision in listening. In W. Hardcastle, & J. Beck (Eds.), *A figure of Speech: a festschrift for John Laver* (pp. 63-91). Mahwah, NJ: Erlbaum.

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

Diehl, R.L., Lotto, A.J., & Holt, L.L. (2004). Speech perception. *Annual Review of Psychology*, *55*, 149-179.

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

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

Eisner, F. & McQueen J.M. (2006) Perceptual learning in speech: Stability over time. *Journal of the Acoustical Society of America*, *119*, 1950-1953.

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

Evans, B.G., & Iverson, P. (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.

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

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

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

Flege, J.E. & Eefting, W. (1987). Cross-language switching in stop consonant perception and production by Dutch speakers of English. *Speech Communication*, *6*, 185-202.

Flege, J.E. & Hillenbrand, J. (1986). Differential use of temporal cues to the /s/-/z/ contrast by native and non-native speakers of English. *Journal of the Acoustical Society of America, 79,* 508-517.

Folstein, J. R., & van Petten, C. (2004). Multidimensional rule, unidimensional rule, and similarity strategies in categorization: event-related brain potential correlates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 1026-1044.

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

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

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

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

Getty, D.J., Swets, J.B., Swets, J.A., Green, D.M.(1979). On the prediction of confusion matrices from similarity judgments. *Perception & Psychophysics*, *26*, 1-19.

Glasberg, B.R., & Moore, B.C.J. (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.

Goldstone, R.L. & Barsalou, L. (1998). Reuniting perception and conception. *Cognition*, *65*, 231-262.

Goldstone, R.L., & Kersten, A. (2003). Concepts and categorization. In A. F. H. R. W. Proctor (Ed.), *Comprehensive handbook of psychology* (Vol. 4: Experimental psychology, pp. 599-621). New Jersey: Wiley.

Gottfried, T.L. & Beddor, P.S. (1988). Perception of temporal and spectral information in French vowels. *Language & Speech*, *31*, 57-75.

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

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

Gureckis, T. M., & Love, B. C. (2003). Human Unsupervised and Supervised Learning as A Quantitative Distinction. *International Journal of Pattern Recognition and Artificial Intelligence*, *17*, 885-901.

Gureckis, T. M. & Love, B. C. (2004). Common Mechanisms in Infant and Adult Category Learning. *Infancy*, 5, 173-198.

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

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

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

Hecaen, H. & Angelergues (1962). Agnosia for faces (proposagnosia). *Archives of Neurology*, *7*, 92-100.

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

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

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

Holt, L.L., Lotto, A. J., & Kluender, K. R. (1998). Incorporating principles of general learning in theories of language acquisition. In M. Gruber, C. Derrick Higgins, K.S. Olson & T. Wysocki (Eds.), *Chicago Linguistic Society, Volume 34: The Panels* (pp 253-268). Chicago: Chicago Linguistic Society.

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

James, W. (1890). *The Principles of Psychology*. New York: Holt.

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

Kawahara, S. (2006). Contextual effects on the perception of duration. *Journal of the Acoustical Society of America, 199,* 3234.

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

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

Kraljic, T., & Samuel, A.G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*, 141-178.

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

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

Kuhl, P.S. (2000). A new view of language acqusition. *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 11850-11857.

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

Ladefoged, P. (1999). "American English." In *Handbook of the International Phonetic Association* (pp. 41-44). Cambridge: Cambridge University Press.

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*, 127-132.

Lively, S.E., Logan, J.S., & Pisoni, D.B (1993).Training Japanese listeners to identify English /r/ and /l/ - 2. The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America*, *94*, 1242-1255.

Lively, S.E., Pisoni, D.B., Yamada, R.A., Tohkura, Y., & Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/ III: Long-term retention of new phonetic categories. *Journal of the Acoustical Society of America*, *96*, 2076-2087.

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

Lotto, A.J. (2000). Language acquisition as complex category formation. *Phonetica*, *57*, 189-196.

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

Love, B.C. (2003). The multifaceted nature of unsupervised category learning *Psychonomic Bulletin & Review*, *10*, 190-197.

Love, B.C., Medin, D.L, and Gureckis, T.M (2004) SUSTAIN: A Network Model of Category Learning. Psy*chological Review*, *11*, 309-332

Lutfi, R. A., Kistler, D. J., Oh, E. L., Wightman, F. L., & Callahan, M. R. (2003). One factor underlies individual differences in auditory informational masking within and across age groups. *Perception & Psychophysics*, *65*, 396-406.

Macmillan, N.A., & Creelman, C.D. (1991). *Detection theory: A user's guide*. New York: Cambridge University Press.

Maddox, W.T. & Ashby, F.G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, *53*, 49–70.

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

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

Maye, J. & Gerken, L. (2000). Learning phoneme categories without minimal pairs. *Proceedings of the 24th Annual Boston University Conference on Language Development*: 522-533.

Maye, J. & Gerken, L. (2001). Learning phonemes: How far can the input take us? In A. H-J. Do, L. Domínguez, & A. Johansen (Eds.), *Proceedings of the 25th Annual Boston University Conference on Language Development* (pp 480-490). Sommerville, MA.

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

Medin, D.L. & Barsalou, L.W. (1987). Categorization processes and categorical perception. In Stevan Harnad (Ed.), *Categorical Perception* (pp. 455-490). Cambridge: Cambridge University Press.

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

Minda, J.P. & Smith, J.D. (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. & Smith, J.D. (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.

Mirman, D., Holt. L.L. & McClelland (2003). Categorization and discrimination of nonspeech sounds: Differences between steady-state and rapidly changing acoustic cues. *Journal of the Acoustical Society of America*, *116*, 1198-1207.

Nearey, T.M. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America*, *101*, 3241-3254.

Nearey, T.M. & Hogan, J.T. (1986). Phonological contrast in experimental phonetics: relating distributions of production data to perceptual categorization curves. In: J.J. Ohala & J.J. Jaeger (Eds.), *Experimental Phonology* (pp 141-161). Orlando, Academic Press.

Nearey, T.M. & Assman, P.F. (1986). Modeling the role of inherent spectral change in vowel identification. *Journal of the Acoustical Society of America*, *80*, 1297-1308.

Norris, D., McQueen, J.M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*, 204-238.

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

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

Pegg, J.E. & Werker, J.F. (1997). Adult and infant perception of two English phones. *Journal of the Acoustical Society of America*, *102*, 3742-3753.

Peperkamp, S., Pettinato, M., & Dupoux, E. (2003). Allophonic variation and the acquisition of phoneme categories. In A. Brown, & F. Conlin (eds.) *Proceedings of the 27th Annual Boston University Conference on Language Development*. (pp. 650-661) Somerville, MA : Cascadilla Press.

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

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

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

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

Polka, L. (1995). Linguistic influences in adult perception of non-native vowel contrasts. *Journal of the Acoustical Society of America*, *97* (2), 1286-1296.

Pomerantz, J.R., & Lockhead, G.R. (1991). Perception of structure: an overview. In G.R. Lockhead and J.R. Pomerantz (Eds.), *The perception of structure* (pp. 1 - 20). Washington, D.C.: American Psychological Association.

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

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

Regehr, G. & Brooks, L.R. (1995). Category organization in free classification: The organizing effect of an array of stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 347-363.

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

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, *4*, 328—350.

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

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

Seger, C.A., Poldrack, R.A., Prabhakaran V, Zhao M, Glover, G.H., & Gabrieli, J.D. (2000). Hemispheric asymmetries and individual differences in visual concept learning as measured by functional MRI. *Neuropsychologia*, *38*, 1316-24.

Shepard, R., Hovland, C., & Jenkins, H. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75*, 1-42.

Shepard, R.N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*, 325-45.

Shepard, R.N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*,1317-1323.

Shepard,R.N. (1962a).The analysis of proximities:Multidimensional scaling with an unknown distance function: Part I. *Psychometrika*, *27*, 125–140.

Shepard,R.N. (1962b).The analysis of proximities:Multidimensional scaling with an unknown distance function: Part II. *Psychometrika*, *27*, 219–246.

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

Smiljanić, R.& Bradlow, A.R. (2005). Production and perception of clear speech in Croatian and English. *Journal of the Acoustical Society of America*, *118*, 1677-1688.

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

Strange, W. (ed.) (1995). *Speech perception and linguistic experience: Issues in cross-language speech research*. Timonium, MD: York Press.

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

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

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

Tversky, A. & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review*, *89*, 123-54.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327-52

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

Werker, J.F. & Lalonde, C.E. (1988). Cross-language speech perception: Initial capabilities and developmental change. *Developmental Psychology*, *24*, 672-683.

Werker, J.F. & Tees, R.C. (1983). Developmental changes across childhood in the perception of non-native speech sounds. *Canadian Journal of Psychology*, *37*, 278-286.

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

Werker, J.F., & Yeung, H.H. (2005). Infant speech perception bootstraps word learning. *Trends in Cognitive Sciences*, *9*, 519-527.

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

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

Appendix A

Sweep rate experiments[15]

15 An altered version of this chapter will be published as Goudbeek, M. & Swingley, D. (2006). Saliency Effects in Distributional Learning. *Proceedings of the Eleventh Australasian International Conference on Speech Science and Technology*. 2006. Auckland, New Zealand.

Introduction

As mentioned in Chapter 3, duration and formant frequency were not the only two auditory dimensions considered for the category learning experiments. This appendix describes experiments with another dimension that was combined with formant frequency, the speed of the rise in frequency of the base frequency of the signal (F0) sweep rate. This dimension was used in a number of unsupervised learning experiments. Primarily because of the difficulty to equalize sweep rate and formant frequency in terms of just noticeable differences, duration was preferred in the experiments presented in the thesis.

First, the same/different pilot experiments that were aimed at equalizing the just noticeable differences for sweep rate and formant frequency are reported. With this just noticeable difference, the range of variation for both dimensions (formant frequency and sweep rate) is set. The just noticeable difference commonly used for formant frequency for this area of perceptual space is 0.12 ERB (Glasberg & Moore). In the pilot experiments a value for sweep rate is that is similar in perceptual saliency is determined.

Second, two unsupervised category learning experiments with these dimensions are reported. Because Experiment 1 showed the dimensions to be far from similar in distinguishabilty/saliency (despite our efforts to create equal just noticeable differences for both dimensions in the pilot), Experiment 2 was conducted to find the

value for sweep rate that had a similar salience as the one for formant frequency in a category learning experiment. In these experiments, the just noticeable difference for sweep rate was systematically varied to find the right just noticeable difference.

Pilot experiments

Method

Subjects

Thirty-seven listeners participated in the pilot experiments. Participants were drawn from the subject pool of the Max Planck Institute for Psycholinguistics and received a small payment for their participation. All were students from the University of Nijmegen and reported normal hearing.

Stimuli

The stimuli were inharmonic tone complexes that were similar to the non-speech sounds used in Chapters 2 and 3. Contrary to those stimuli, they did not differ in formant frequency and duration, but in formant frequency and sweep rate. Sweep rate is defined in octaves per second as the speed with which the first formant rises with time. The sweep values of the stimuli ranged between 2 octaves per second and 15 octaves per second, depending on condition (See table A1). The ERB rates (Glasberg & Moore, 1990) of the different conditions are also presented in table A1.

Table A1.
Stimulus characteristics per condition and the experimental properties of the different conditions.

Condition	Min	Max	Stepsize	Tested differences	Trials	High ERB	Low ERB
1	2.2 oct/s	2.8 oct/s	0.2 oct/s	0.2 / 0.4 / 0.6 oct/s	192	17.9 ERB	20.6 ERB
2	5.0 oct/s	15 oct/s	1.0 oct/s	1.0 / 2.0 oct/s	400	18.8 ERB	19.7 ERB
3	5.0 oct/s	15 oct/s	0.5 oct/s	0.5 / 1.5 oct/s	400	18.8 ERB	19.7 ERB

Procedure

All conditions consisted of same/different judgment tasks in which half of the stimulus pairs were same trials and half were different trials. Listeners were seated comfortably in an experiment room and listened to stimulus pairs over Sennheiser headphones (HD 270). If they considered the sounds to be the same, they pressed a button labeled with (the Dutch equivalent of) "same". If they considered the sounds to be different, they pressed a button labeled with (the Dutch equivalent of) "different". All conditions lasted for about 30 minutes and participants were given the possibility of a break halfway through the experiment. The comparisons were done at multiple levels of ERB and sweep rate. For example, in Condition 1, the difference between 2.2 octaves per second and 2.4 octaves per second was compared at different frequencies (ERB levels). This way, possible interactions between the two dimensions could be investigated. Differences in sweep rate were also compared at different levels to investigate possible differences in just noticeable differences at different levels of sweep rate. For example, in Condition 2 the differences between 5.0 and 6.0 octaves per second and that between 14.0 and 15.0 octaves per second were compared.

Results

All three conditions of the same/different pilot experiment yielded hit rates and false alarm rates that were used to compute the d' values associated with each difference in sweep rate. As a d' of about 1 is considered to reflect two perceptually separable stimuli, the goal of the pilot experiment is to find a sweep rate with a d' as close to 1 as possible. Table A2 shows the results of the pilot experiment.

Table A2.
Mean d' values and their standard deviations of the differences tested in all three conditions.

		Condition 1 (N=13)			Condition 2 (N=14)		Condition 3(N=10)	
		Difference (oct/s)						
	ERB level	0.2	0.4	0.6	1.0	2.0	0.5	1.5
d' (σ)	Low	0.28 (0.68)	0.85 (0.86)	1.09 (0.83)	1.93 (0,85)	3.12 (0.86)	0.58 (0.79)	1.28 (0.86)
	High	0.49 (0.51)	0.69 (0.82)	1.26 (0,82)	1.46 (0.32)	2.93 (0.89)	0.35 (0.47)	1.43 (0.98)

According to the data presented in table A2, a difference sweep rate between 0.6 and 1.5 octaves per second has a d' of approximately 1. Condition 1 showed that sweep rate differences lower than 0.4 octave per second were difficult to distinguish. Condition 2 and 3 showed that sweep rates higher than 1.5 were very easy to distinguish. Somewhere between 0.6 and 1.4 lies the d' value of 1 looked for in this pilot experiment. Because the d's for 0.5 from Condition 3 were considerably lower than 1, we decided upon a sweep rate of 1.0 octave per second to constitute a just noticeable difference in the following category learning experiments.

Table A2 appears to indicate that a high ERB level is associated with higher mean d' values for the sweep rate differences. However, the d's of the different ERB level do not differ significantly (all $p > 0.18$, $t[max] = 0.95$). This absence of a significant

difference between the higher and lower ERB rate justifies the use of one just noticeable difference for sweep rate at all ERB levels in the categorization experiments. In other words, there is no evidence of interaction between the two dimensions in this part of perceptual space.

Categorization experiments

Two experiments are presented here. The first experiment uses the just noticeable difference for sweep rate determined in the pilot experiments (1.0 octave per second) and the just noticeable difference for formant frequency (0.12 ERB) derived from Glasberg & Moore (1990). The results show that the just noticeable difference for sweep rate was not comparable to that for formant frequency. In Experiment 2 the sweep rate is systematically varied to equalize the distinguishabilty of both dimensions.

Experiment 1

Method

Subjects

Thirty-six students from the MPI subject pool participated in the experiment. All were students at the University of Nijmegen and participated in return for a small payment. None reported any history hearing difficulties.

Stimuli

The 224 learning stimuli (2 categories x 112 stimuli in each category) were inharmonic sound complexes that differed in both formant frequency and sweep.

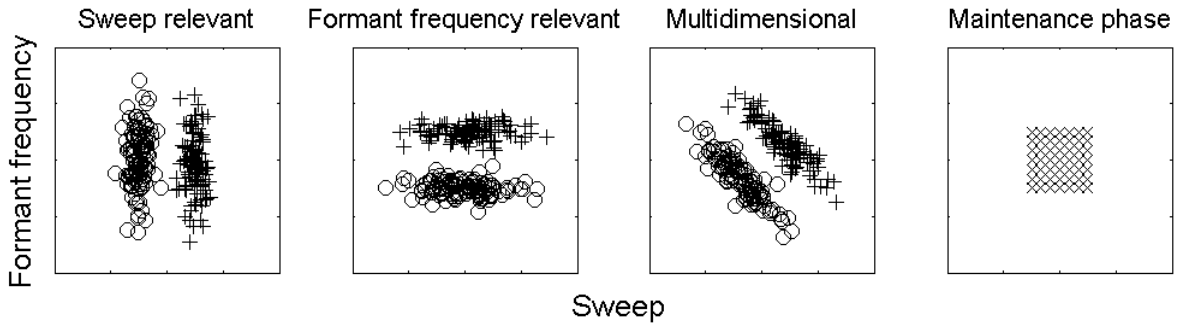

Figure A1. The basic experimental design of the experiments in this chapter: training phases with distributional information (either one or two relevant dimensions) and a neutral maintenance phase.

The probability distributions of the stimuli identified the relevant and irrelevant dimension for the listeners (see Figure A1). In the learning phase of Conditions 1 through 4, formant frequency was the relevant dimension (see the second panel of Figure A1) whereas in the learning phase of in Conditions 5 and 6, sweep rate was the relevant dimension (see the first panel of Figure A1). In Condition 7, both dimensions were relevant (see the third panel of Figure A1). The 49 stimuli of the maintenance phase of each condition were positioned in an equidistant (7 x 7) grid, thus not providing any distributional information to the listeners (see the fourth panel of Figure A1). The labeling sounds were two simple and easily distinguishable sounds.

Design

All conditions had a similar design with a learning phase and a maintenance phase (see Figure A1). In the learning phase, listeners listened to stimuli under a number of

conditions but did not categorize them. In the maintenance phase, they had to categorize the stimuli as they saw fit. Depending on condition, either formant frequency or sweep rate was the relevant dimension of variation in the learning phase.

Because we wanted to investigate learning under several conditions, the conditions also differed in the task subjects had to perform in the learning phase. First, we wanted to compare learning without any supervision (the listening condition) with learning where there is a perfectly correlated cue to category membership (the labeling condition). This perfectly correlated cue was an auditory label that directly followed the sound before it. To equalize all conditions in terms of auditory complexity, there was an *uninformative* label between the stimuli in the other conditions.

Second, to make it harder for our listeners to attend to the distributional information in the stimuli, we added a condition with a lexical decision task to the experiment. This task was combined with the uninformative labels (lexical decision only condition), and with the informative labels (the labeling and lexical decision condition).

This results in the following seven conditions: Condition 1: *Listening* with intermittent *uninformative* sounds, formant frequency relevant; Condition 2: *labeling* with intermittent informative labels, formant frequency relevant; Condition 3: *listening* with intermittent *uninformative* sounds and a *lexical decision* task, formant frequency relevant; Condition 4: *labeling* with informative labels and a *lexical decision* task, formant frequency relevant; Condition 5: *listening* with intermittent *uninformative* sounds, sweep rate relevant; Condition 6: *labeling* with intermittent

informative labels, sweep rate relevant; Condition 7: *listening* with intermittent *uninformative* sounds, both formant frequency and sweep rate relevant.

Procedure

Listeners were seated in a soundproof booth and listened to the stimuli over Sennheiser headphones (HD 270). In the learning phase, they listened passively to the stimuli and sometimes had, depending on condition, another task or another source of information besides the distributional information in the stimuli. Each learning phase of each condition contained 448 stimuli (112 stimuli x 2 categories x 2 repetitions) and was interrupted by a pause after 224 stimuli.

After the learning phase, listeners entered the maintenance phase where they had to categorize 196 (49 stimuli x 4 repetitions) maintenance stimuli as they saw fit. The maintenance phase was intended to neutrally scan the categorization tendencies of the listeners without providing new information about the category distributions.

Results

Since listeners only respond in the maintenance phase, this is the only phase that can be analyzed. To probe for possible changes in categorization strategies during the maintenance phase, the maintenance phase was analyzed in two parts.

For a binary choice problem with two categories, an analysis using logistic regression is the analysis of choice. A logistic regression analysis yields a β-weight for each predictor entered into the analysis. In this case, the dimensions were entered as predictors for the categorization response. Table A3 lists the mean β-weights of each dimension in each phase of each condition.

Table A3.
Mean β-weight and their standard deviations of the first and second part of the maintenance phase for all seven conditions.

		Maintenance phase 1				Maintenance phase 2			
Condition	Relevant dimension	$\mu\beta_{sweep}$	σ	$\mu\beta_{freq}$	σ	$\mu\beta_{sweep}$	σ	$\mu\beta_{freq}$	σ
Listening	Formant frequency	1.58	0.44	0.31	0.20	1.43	0.21	0.27	0.14
Labeling	Formant frequency	1.38	0.52	0.33	0.15	2.02	0.82	0.43	0.31
Lexical decision	Formant frequency	1.33	0.89	0.44	0.36	1.74	0.65	0.39	0.21
Labeling and LD	Formant frequency	1.17	0.59	0.39	0.12	1.30	0.19	0.31	0.19
Listening	Sweep rate	1.74	0.94	0.44	1.24	1.58	0.43	0.39	0.23
Labeling	Sweep rate	1.75	0.94	0.44	0.24	1.59	0.44	0.40	0.23
Listening	Both	1.30	0.69	0.52	0.29	1.38	0.64	0.45	0.29

The data presented in Table A3 show that the listeners used sweep rate much more in their categorization compared to formant frequency, irrespective of condition or whether it was the relevant dimension or not. An ANOVA with Dimension (relevant versus irrelevant) and Part of the maintenance phase (first versus second part) as within-subject variables and Condition and Orientation (formant frequency relevant, sweep rate relevant, or both relevant) as between-subject variables and the β-weights for both dimensions as dependent variables showed no significant main effects of Part, Dimension, Orientation or Condition (all $F\,[1,45] < 1$, *n.s.*).

Because the above design is not perfectly balanced, the effect of the relevance of the dimensions and the orientation of the distributions and the conditions was further investigated by concentrating on conditions 1, 2, 5, and 6 (see Figure A2).

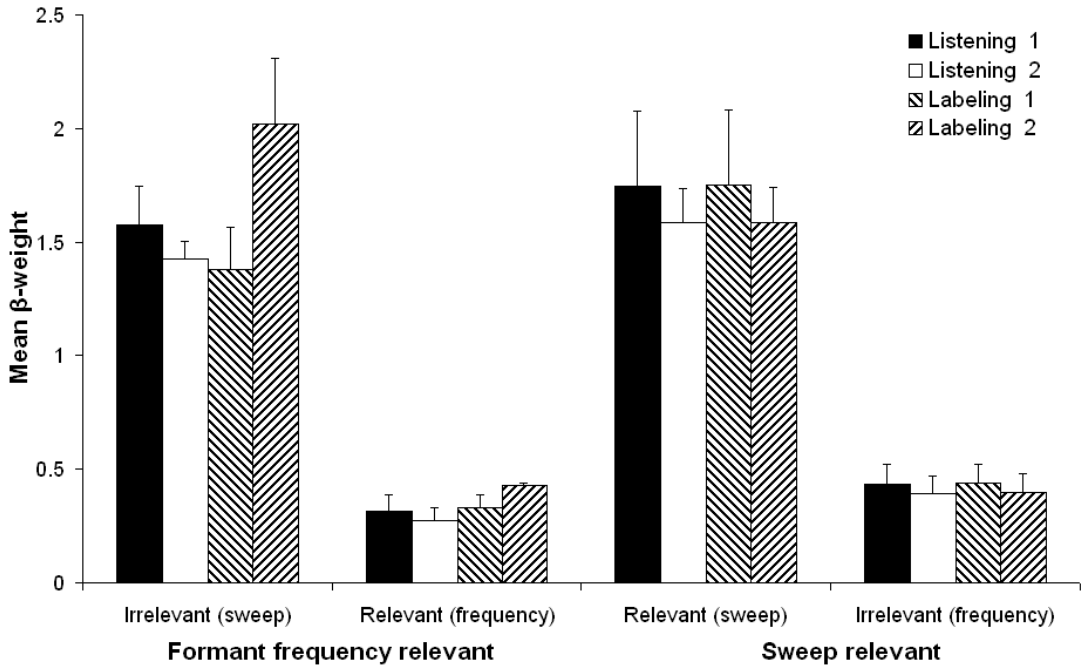

Figure A2. Mean β-weights of the first and second part of the maintenance phases of Conditions 1, 2, 5, and 6; the listening and labeling conditions with one relevant dimension of variation, either formant frequency (Conditions 1 and 2) or sweep rate (Conditions 5 and 6). Vertical error bars represent one standard error. Note that sweep is the dimension used irrespective of whether it is the relevant dimension.

With these conditions, we performed an ANOVA with Part of the maintenance phase (first versus second part) and Dimension (relevant versus irrelevant) as within-subjects variables and Orientation (formant frequency relevant versus sweep rate relevant) and Condition (listening versus labeling) as between-subjects variables. There were no significant effects for Dimension ($F [1,27] =0.004$, *n.s.*) or Orientation ($F [1,27] = 0.46$, *n.s.*) showing that listeners were not sensitive to the different category structures. The interaction between Orientation and Dimension was highly significant ($F [1,27] = 139.71$, $p < 0.000$) indicating the preference for sweep rate, irrespective of whether it was the relevant condition or not.

Because of the lack of significant effects of the experimental manipulations, the chosen just noticeable difference for sweep rate was reconsidered. Just noticeable differences determined with a same/different paradigm apparently do not transfer to a categorization experiment. In Experiment 2, we systematically manipulated the size of the difference in sweep rate in an attempt to find a just noticeable difference for sweep rate that was equal to that chosen for formant frequency in a category learning experiment.

Experiment 2[16]

Method

The rationale of Experiment 2 is that it must be possible by systematically varying the differences in sweep rate, to find the sweep rate just noticeable difference that is equal to the just noticeable difference used for formant frequency (0.12 ERB).

Subjects

Twenty-four participants (four in each condition) were drawn from the MPI subject pool and took part in the experiment. All were students at the University of Nijmegen and received a small payment for their contribution. None reported any hearing difficulties.

16 Marloes van der Goot and Maarten Jansonius are thanked for their help in recruiting the participants and running the experiments.

Stimuli

The stimuli were identical to those used in Experiment 1: inharmonic sounds that differed in formant frequency and sweep rate. Depending on condition, either the variation in formant frequency was relevant for distinguishing the categories and sweep rate was irrelevant or vice versa (see the first and second panel of Figure A1). The just noticeable difference for sweep rate was also varied with condition: conditions with "Sweep 2" had a sweep rate of 0.5 octave per second; conditions with "Sweep 4" had a sweep rate of 0.25 octave per second and conditions with "Sweep 8" had a sweep rate of 0.125 octave per second.

Procedure

The procedure was identical to the labeling conditions of Experiment 1. In the learning phase, listeners heard a stimulus that was immediately followed by an acoustical label that correlated perfectly with category membership. In the maintenance phase, listeners were asked to categorize the stimuli as they saw fit.

There were six experimental conditions (2 category structures x 3 sweep rate levels) in the experiment. Four listeners participated in each condition.

Results

The results from the maintenance phase were again analyzed with a logistic regression analysis yielding a β-weight indicating each subject's use of each dimension. Table A4 and Figure A3 show the mean β-weights for all six conditions. When the just noticeable difference for sweep rate was set at 0.5 octave per second,

listeners still had a higher β-weight for sweep rate, irrespective of whether it was the relevant dimension or not.

When the just noticeable difference for sweep rate was set to 0.125 octave per second, however, the variation in sweep rate was too small and the β-weight for formant frequency was higher than that for sweep rate. Again, this was independent of whether formant frequency was the relevant dimension or not.

Table A4.
Mean β-weight for the two dimensions (formant frequency and sweep rate) for all three levels of sweep rate (0.5 octave per second, 0.25 octave per second, and 0.125 octave per second) and the two category orientations (formant frequency relevant and sweep rate relevant).

	Formant frequency relevant			Sweep rate relevant		
	Sweep 2	Sweep 4	Sweep 8	Sweep 2	Sweep 4	Sweep 8
			Maintenance Phase 1			
	β (σ)	β (σ)	β (σ)	β (σ)	β (σ)	β (σ)
Formant frequency	0.47 (0.27)	1.35 (0.38)	1.38 (0.27)	0.29 (0.13)	0.85 (0.40)	1.41 (0.60)
Sweep	1.12 (0.42)	0.37 (0.27)	0.22 (0.23)	1.38 (1.31)	0.97 (0.48)	0.29 (0.30)
			Maintenance Phase 2			
	β (σ)	β (σ)	β (σ)	β (σ)	β (σ)	β (σ)
Formant frequency	0.33 (0.13)	1.81 (0.68)	1.66 (0.52)	0.26 (0.22)	1.01 (0.24)	1.18 (0.89)
Sweep	1.75 (0.48)	0.51 (0.12)	0.27 (0.08)	1.36 (0.18)	1.21 (0.86)	0.08 (0.05)

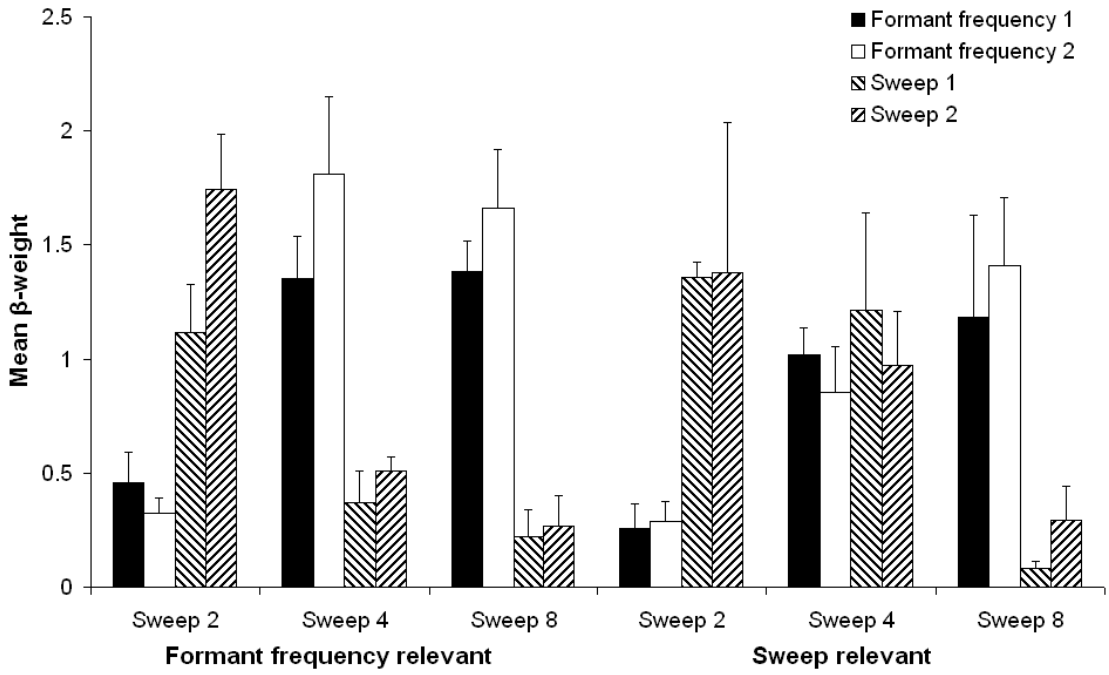

Figure A3. The mean β-weight per orientation (formant frequency relevant or sweep rate relevant) and per level of sweep rate (0.5 octave per second; 0.25 octave per second; 0.125 octave per second) for the first and second part of the maintenance phase.

Finally, when the just noticeable difference for sweep rate was 0.25 octave per second, the experimental manipulations are not washed out by the differences in salience of the different dimensions. With a sweep rate of 0.25 octave per second the relevant dimension is the one that is used most by listeners. The effect is still quite small when sweep rate is the relevant dimension, but compared to when formant frequency is relevant, the differences are considerable. An ANOVA with Part of the maintenance phase (phase 1 versus phase 2) and Dimension (relevant versus irrelevant) as within-subjects variables and Orientation (formant frequency relevant versus sweep rate relevant) as between-subjects variable and the β-weights of each dimension a dependent variables indicated a marginally significant main effect of

Dimension ($F [1,6] = 3.99$, $p < 0.09$). This shows that listeners were able to determine and use the relevant dimension in their categorizations.

In sum, a sweep rate of 0.25 octave per second is the best counterpart for a just noticeable difference of 0.12 ERB. Because the β-weights for formant frequency are still considerably higher than those for sweep rate, a sweep rate between 0.25 octave per second and 0.5 octave per second (0.375 octave per second) might represent an even more similar just noticeable difference.

Conclusion

These experiments have shown the difficulty of equalizing the just noticeable differences for two important dimensions in speech recognition; formant frequency and sweep rate. Even after careful piloting with same/different experiments, sweep rate was still the dominant dimension in Experiment 1. The results of Experiment 2 show, however, the validity of the logic of conducting consecutive categorization experiment with differing sweep rates to find the sweep rate that best fits the just noticeable difference for formant frequency.

Appendix B

Unidimensional category learning

Introduction

Most distinctions between speech sounds are multidimensional in nature. However, there are examples of one dimension accounting for most of the difference between two speech sounds. This appendix presents experiments investigating the learning of a unidimensional distinction under several conditions. Contrary to the stimuli presented in the rest of the dissertation, the stimuli used in this appendix vary only in one dimension (spectral peak) and do not have an irrelevant dimension of variation. The conditions under which learning of these unidimensionally separable categories is investigated are based on reasoning put forward before in this thesis (see Chapters 2, 3 and 4), but will be briefly repeated below.

Consider the infant's situation: In a cacophony of sounds, only some are language. The infant's task is to extract the relevant patterns in this input. A striking aspect of infant auditory category learning is the absence of explicit supervision. When infants learn categories, there is no observable behavior so acquisition must have taken place without explicit supervision. Also, this learning certainly cannot be verbally mediated; it has to be implicit.

A way to account for the exceptional performance of infants in the absence of supervision or verbal mediation is to consider them as statistical pattern recognizers. They perform a distributional analysis of the incoming acoustic data. This way, they

learn the categorical regularities in the input. Adult learning of auditory categories, however, can probably also be supervised and verbally mediated.

The experiments presented here examined auditory category learning in adults, comparing conditions which encouraged implicit learning (like that in infants) with conditions which encouraged explicit learning by manipulating the presence or absence of a secondary lexical decision task. When present, the lexical decision task ought to prevent or at least hinder explicit learning. The lexical decision task had subjects decide on words and nonwords (both with a 50% probability) that were presented together with the category exemplars (with varying stimulus onset asynchrony). We also manipulated the presence or absence of feedback in the form of a perfectly correlated auditory cue (the *label*). In order to equalize the conditions for auditory complexity and the possible effect of backward masking in counteracting the facilitative effect of the auditory labels, the stimuli in conditions without the informative auditory label were followed by a label that was *not* informative of the category of the preceding stimulus.

Additionally, we wanted to investigate the a priori categorization tendencies of listeners and the speed of learning. We did this by either removing the learning phase or drastically shortening it.

The first four conditions followed a 2 x 2 design with two independent variables: the presence or absence of supervision (implemented as the presence of a perfectly correlated auditory cue 300 ms after the stimulus), and whether learning was explicit or implicit (manipulated by the secondary task).

This created the following four conditions (identical to those in Appendix A). Condition 1, where 440 learning stimuli were presented with *uninformative* labels and listeners had to rely solely on the distributional information in the stimuli.

Condition 2, where the labels were *informative* about category membership of the preceding stimulus and listeners thus had two sources of information. Condition 3, in which participants had to perform a secondary task (a go/nogo lexical decision task with words and nonwords between the learning stimuli) and the stimuli were followed by *uninformative* labels. Finally, Condition 4 where both the *informative* labels were present and listeners had to perform the lexical decision task.

Two additional conditions investigated listeners' initial categorization tendencies and the speed of learning. In Condition 5, subjects categorized the stimuli as they saw fit, without a preceding learning phase. Condition 6 was identical to Condition 3 (a secondary task without informative labels) but listeners were tested after a quarter of the learning phase (110 stimuli instead of the usual 440).

Table B1. Properties of all six experimental conditions.

Condition	Supervision	Distraction	Learning phase	N
1	No	No	440 stimuli	8
2	Yes	No	440 stimuli	9
3	No	Yes	440 stimuli	9
4	Yes	Yes	440 stimuli	14
5	No	No	No	9
6	No	Yes	110 stimuli	14

Method

Subjects

All 63 participants were students of the University of Nijmegen and were drawn from the MPI subject pool. All reported normal hearing. Subjects were randomly assigned to the conditions.

Stimuli

Two auditory categories were constructed. They were complex inharmonic patterns that varied in the frequency of their spectral peak. Table B2 lists the mean spectral peak, the standard deviation and the range of the stimulus dimension of the two categories.

Table B2.
Stimulus characteristics of the learning stimuli.

Category	μ. (Hz/ERB)	σ (Hz/ERB)	Range (Hz/ERB)
A	1291.2 / 17.6	92.5 / 3.2	1078.7 - 1531.6 / 16.2 - 19.0
B	1744.8 / 20.0	120.2 / 3.9	1468.9 - 2057.0 / 18.6 - 21.4

Each category contained 110 different stimuli. To make category learning easy, there were only eight ambiguous (overlapping) items (four per category) and 102 unambiguous ones. The eleven maintenance phase stimuli ranged between the means of the learning phase stimuli in equal steps and thus containing no distributional information that could be of assistance in categorization. Figure B1 displays the categories' probability density functions and the range of the maintenance phase (the dashed line).

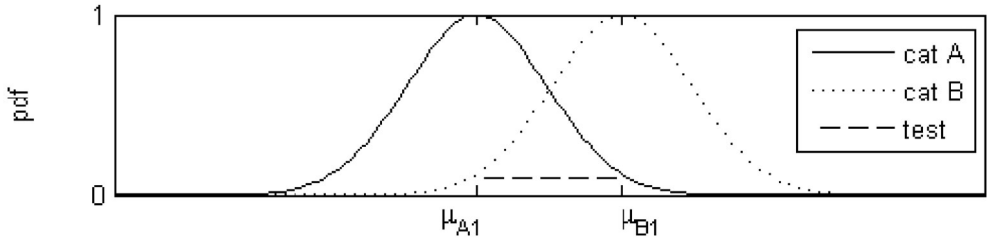

Figure B1. The probability density functions of the two learning categories and the maintenance phase.

Procedure

Listeners were seated in a soundproof booth and listened to the stimuli over Sennheiser headphones (HD 270). In the learning phase, they listened passively to the stimuli and sometimes had, depending on condition, had another task or another source of information besides the distributional information in the stimuli. Each learning phase of each condition contained 440 stimuli (110 stimuli x 2 categories x 2 repetitions) and was interrupted by a pause after 220 stimuli. They received feedback on their lexical decision judgments in the pause.

After the learning phase, listeners entered the maintenance phase where they had to categorize 220 maintenance stimuli (11 stimuli x 20 repetitions). The maintenance phase was intended to neutrally scan the categorization tendencies of the listeners without providing new information about the category distributions.

Results

Category judgments were tested with 11 stimuli evenly spaced between the means of the categories. These categorical responses on continuous stimuli are best analyzed using a logistic regression technique (Agresti, 1990). The β-weights yielded by this analysis indicate to what extent the variation in the stimuli was used by the listeners in their category judgments. Figure B2 displays the mean β weights for the six conditions.

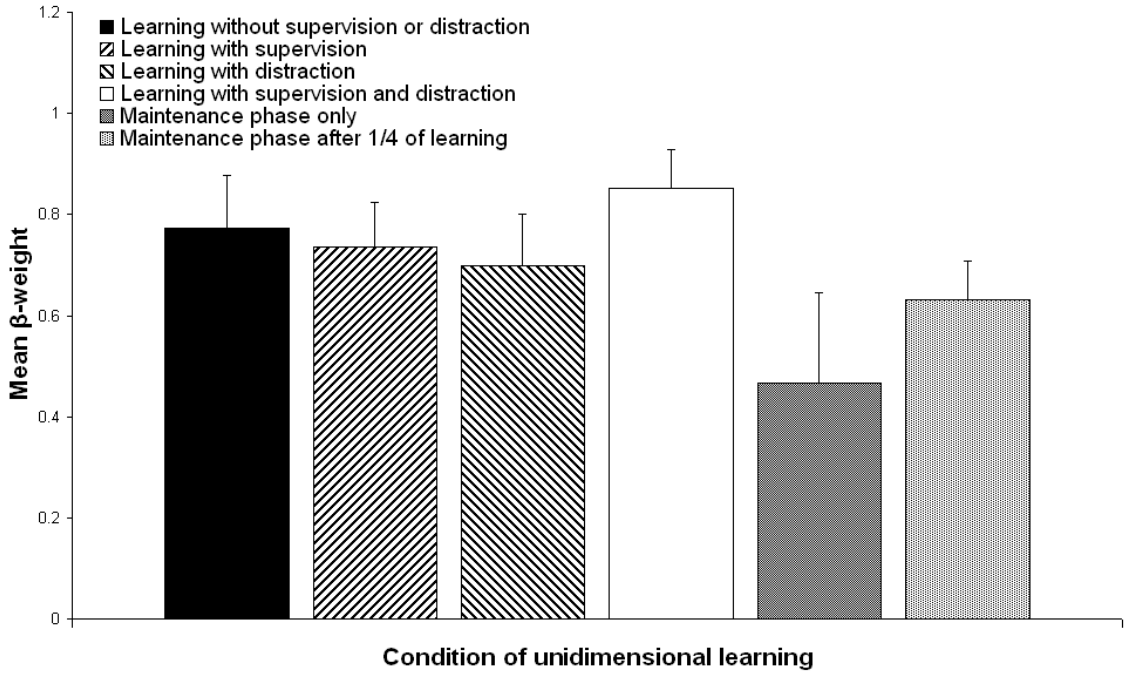

Figure B2. Mean β-weights for the six experimental conditions. Error bars indicate standard errors.

Figure B2 shows that the conditions did not differentiate much. Performance was only negatively affected when listeners did not enter a learning phase before categorizing the stimuli (Condition 5). The effects of supervision, distraction and the learning phase were statistically evaluated in an ANOVA with the β-weights as dependent variable and Supervision, Distraction and the presence of a Learning phase as independent (between subjects) variables. There were no significant main effects of either Supervision ($F [1,48] = 0{,}24$, *n.s.*) or Distraction ($F [1,48] = 0{,}031$, *n.s.*), nor was there a significant interaction between these two ($F [1,48] = 0{,}65$, *n.s.*). The comparison of all conditions with a learning phase with the one without training did show a marginally significant difference in favor of the conditions with a learning phase ($F [1,48] = 3{,}12$, $p < 0{,}08$). The comparison between a learning phase of 440 stimuli with distraction (Condition 3) with a learning phase of 110 stimuli with

distraction (Condition 6) showed no significant difference (F [1,27] = 0,31, *n.s.*). Listeners apparently pick up the distributional information very quickly.

Discussion

The results showed that learning of unidimensionally varying auditory categories was possible with and without (implicit) supervision by a perfectly correlated auditory cue. Even the presence of a distracting task did not hamper performance. These findings are surprising. There are a number of explanations for these results. First, it could be that the category structure was exceptionally easy to learn, given the small amount of overlap between the two categories. The poor performance of subjects who were tested *without* previous learning makes this interpretation less likely. Second, the amount and speed of learning was perhaps so great that differences between conditions faded away. The absence of a significant difference in performance after a learning face of 220 stimuli and a learning phase of 55 stimuli seems to favor this interpretation. Listeners learned to categorize some unidimensional stimuli very quickly.

Appendix C

Improving unsupervised category learning

Introduction

Multidimensional unsupervised category learning is very fragile. Categorization rules using two dimensions that are acquired in the learning phase often get lost quickly in the maintenance phase. For example, in Chapter 3 listeners learn to use a multidimensional categorization rule in the learning phase, but revert to using only one (either duration or formant frequency) instead of two dimensions in the maintenance phase. Appendix A shows that the performance of listeners is highly dependent on the distributional properties of the stimuli. This appendix presents two series of experiments that tried to improve categorization performance in the maintenance phase of multidimensional category structures, i.e., category distinctions where both dimensions are relevant. This was done by manipulating the distributional characteristics of the stimuli (range and standard deviation) or by changing the procedure in the learning phase and the experimental instructions. Figure C1 shows the general idea behind both experiments: unidimensional categorization rules are ineffective in multidimensional categorization because they lead to many incorrect categorizations. If the costs of a unidimensional rule are so high, then why do listeners not maintain their learned multidimensional categorization rule? And how can they be brought to do so?

Both experiments attempt to direct listeners away from unidimensional rules where listeners make a lot of categorization errors (the left and middle plane of

Figure C1) towards a multidimensional rule where listeners theoretically make no errors (the right plane of Figure C1). In Experiment 1 the distributional properties of the categories are manipulated in order to achieve this, whereas in Experiment 2 the instructions and the information given to the listeners is manipulated to achieve better (i.e., more multidimensional) performance.

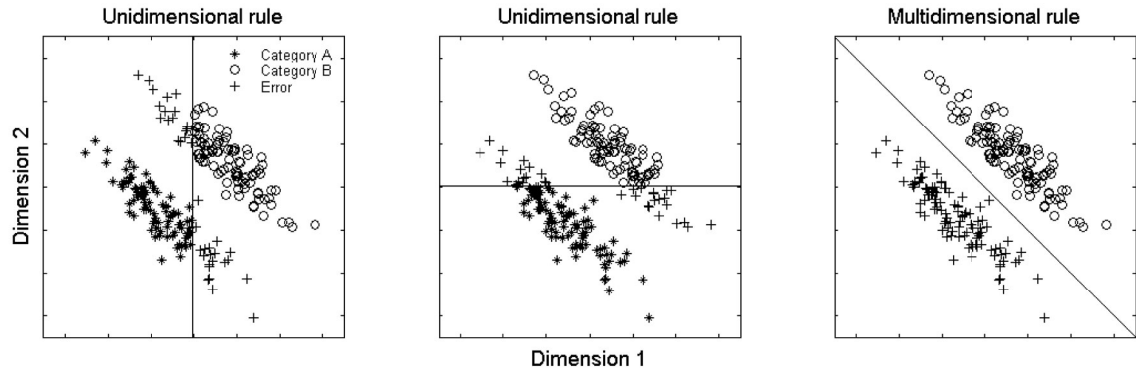

Figure C1. Correct and incorrect categorizations with unidimensional and multidimensional categorization rules.

Experiment 1[17]

In Experiment 1 the category structure was manipulated in order to improve the performance of the participants in the maintenance phase. More precisely, the distance between the means as well as the standard deviations were manipulated (see Stimuli section).

17 Experiment 1 was carried out with financial support from the Dutch Scientific Council. We further thank Keith Kluender, University of Wisconsin, Madison for financial and other assistance with these experiments.

Method

Subjects

All 30 participants were psychology students from the University of Wisconsin, Madison. They received course credit for their participations and gave informed consent before taking part in the experiment. The number of participants was 12, 11 and 7 in Condition 1, 2, and 3 respectively.

Stimuli

All experiments consisted of a learning and a maintenance phase. The 224 learning stimuli (112 in each category) were inharmonic complex sounds differing in duration and their spectral peak (See Chapter 2 for details concerning stimulus construction). As mentioned, the conditions differed in the extent to which they encouraged multidimensional learning by differences in their range and standard deviations. Figure C2 shows the stimulus distributions of the three conditions and Table C1 lists the theoretically optimal percentages correct for .each condition.

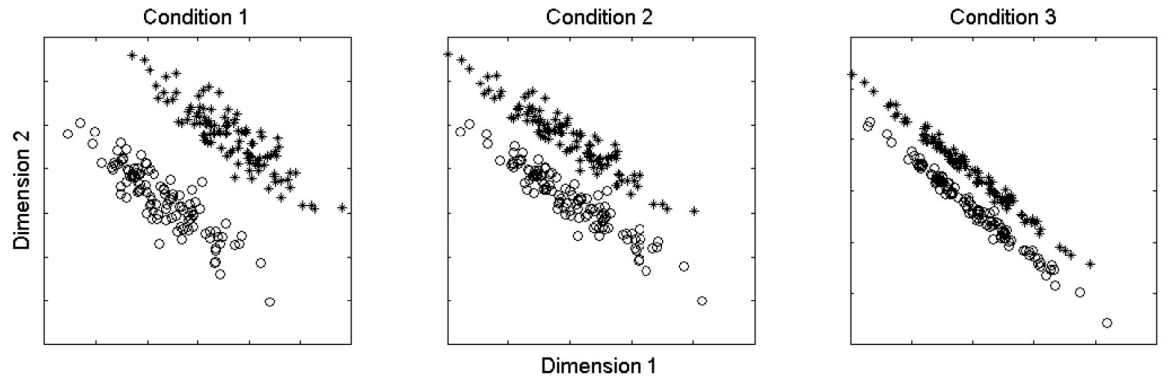

Figure C2. Category structures of the three conditions of Experiment 1. The manipulation of the distributional properties is intended to encourage distributional learning by increasing the difference in optimal percentage correct between unidimensional and multidimensional categorization rules.

Table C1.
Optimal percentages correct for all three conditions for the unidimensional and the multidimensional categorization rules.

Condition	Unidimensional	Multidimensional
1	82%	100%
2	70%	100%
3	57%	100%

As Figure C2 and Table C1 show, a multidimensional categorization rule is always beneficial, but the difference with a unidimensional rule differs depending on the distributional properties of the stimuli.

The maintenance stimuli were generated by constructing an equidistant grid of 49 stimuli (7 by 7) in the same perceptual space as the learning stimuli. For details regarding the maintenance stimuli, see Chapter 2. All stimuli were RMS matched to ensure a constant sound pressure level at the headphones of 65 dB.

Procedure

The listeners were placed in a soundproof booth. After they had received written instructions, they pressed a button to start the experiment. They were instructed to assign sounds to one of two buttons. If the sound was correctly assigned, a light above the button would light up. If not, the light belonging to the other button turned on. With this feedback, the listeners were asked to correctly classify as many stimuli as they could.

Each stimulus was presented in two randomized blocks, resulting in 448 trials (112 stimuli x 2 categories x 2 blocks). This learning phase lasted about 25 minutes, depending on the response speed of the listeners. In condition 3, listeners received an additional block, resulting in 672 (112 x 2 x 3) trials.

After the learning phase, listeners entered the maintenance phase where they categorized the maintenance stimuli as they say fit by pressing the appropriate button. In the maintenance phase, feedback as well as distributional information was absent.

Results

As in the multidimensional experiments in Chapters 2 through 4, the results were analyzed using a polar transformation of the β-weights of the logistic regression analysis, resulting in the angle ϕ (with $\frac{1}{2}\pi$ indicating unidimensional use of formant frequency, 0 indicating unidimensional use of duration and $\frac{1}{4}\pi$ indicating multidimensional categorization) and the distance (consistency) measure A.

Table C2.
The results of Experiment 1, the mean Φ and A values as well as their standard deviations of the learning and the maintenance phase of the three conditions. Additionally, the number of listeners using both dimensions significantly ($p < 0.05$) in their categorization is shown.

Condition	N		Learning	N_{multi}	Maintenance	N_{multi}
1	12	Φ (σ)	0.27 (0.03)	12	0.12 (0.20)	3
		A (σ)	1.44 (0.57)		1.27 (0.41)	
2	11	Φ (σ)	0.29 (0.04)	11	0.05 (0.20)	3
		A (σ)	1.24 (0.57)		1.06 (0.79)	
3	7	Φ (σ)	0.11 (0.29)	4	0.00 (0.26)	0
		A (σ)	0.29 (0.23)		0.79 (0.63)	

The results shown in Table C2 indicate that performance in the learning phases of Condition 1 and 2 was good. The mean value of the angle ϕ was close to $\frac{1}{4}\pi$ indicating multidimensional performance and A was quite large. In the maintenance phase, however, multidimensional categorization was all but absent with ϕ close to zero indicating the often observed preference for duration. The relatively large value

for A shows that subjects used their (incorrect) unidimensional rule consistently. The low A and the small number of listeners using both dimensions in Condition 3, suggest that the distributional manipulations in Condition 3 did more harm than good. The improved performance in A in the maintenance phase indicated the sensitivity of listeners to the (absence of) distributional information and feedback, as they were better able to maintain an (incorrect) unidimensional categorization rule.

Statistical evaluation of ϕ confirmed the impression that listeners were unable to maintain a multidimensional categorization rule. In the learning phases of Condition 1 and 2, Φ differed significantly from both 0 and $\frac{1}{2}\pi$ ($t_{min} = 18.1$, $p < 0.05$) indicating multidimensional categorization. In the maintenance phase ϕ did not differ significantly from 0 (and very significantly from $\frac{1}{2}\pi$ ($t_{min} = 6.4$, $p < 0.05$) showing the overall preference for a unidimensional categorization rule with duration as the relevant dimension.

After assuring all A's differed significantly from zero in all conditions ($t_{min} = 3.3$, $p < 0.05$) an ANOVA with A as a dependent measure and condition as the independent variable was conducted. This showed the conditions to differ significantly in the learning phase ($F [2,27] = 11.7$, $p < 0.05$). but not in the maintenance phase ($F [2,27] = 1.36$, n.s.). Post hoc tests (Tukey HSD) on the learning phase showed Condition 3 to differ significantly from the other two.

Discussion

The goal of Experiment 1 was to increase the number of listeners that used a multidimensional categorization rule in the maintenance phase by manipulating the distributional properties of the categories. In effect, the condition that punished

unidimensional categorization the most, Condition 3, had the worst multidimensional performance. Apparently, listeners do not respond to these properties as predicted. An explanation for this result might be that the categories in Condition 3 were too difficult to separate for the listeners. The diagonal distance between the means may have simple been too small to be perceptually separable. Another possible explanation for this finding could be that there is so much negative reinforcement (the error rate for an initial unidimensional rule is very high) that listeners give up before they can discover the multidimensional rule.

The distributional properties of the experiments presented in Chapter 2 to 4 were based on the results obtained in this experiment. Since the manipulations had no positive effects on performance and were, in their extreme form, even disadvantageous, the distributional properties of Condition 1 were used.

Experiment 2

Experiment 2 tried to improve multidimensional categorization performance in the maintenance phase by manipulating the feedback listeners receive in the learning phase and by giving the listeners explicit instructions regarding their categorizations (Condition 3).

Method

Subjects

Eighteen students of the University of Nijmegen participated in the experiment in return for a small payment. All were drawn from the MPI subject pool and reported normal hearing.

Stimuli

The stimuli were identical to those in Condition 1 of Experiment 1.

Procedure

All conditions had a learning phase and a maintenance phase. The procedure of the maintenance phase was identical in all three conditions: listeners were asked to categorize the well-known equidistantly spaced grid as they saw fit. The learning phase differed according to condition. In Condition 1, listeners received trial-by-trial right/wrong feedback on their responses. In Condition 2, listeners received right/wrong feedback on their responses and were provided with perceptual anchors (consisting of the means of the categories) for the first 40 trials to aid in their

categorization. In Condition 3, listeners received a written explanation of the distributional properties of the stimuli (explaining the diagonal categorization rule and the importance of integrate the two stimulus dimensions to avoid errors), additional to the anchors and the feedback.

Results

Table C3.
The results of Experiment 2, the mean ϕ and A values as well as their standard deviations of the learning and the maintenance phase of the three conditions. Additionally, the number of listeners using both dimensions significantly ($p < 0.05$) in their categorization is shown.

Condition	N		Learning	N_{multi}	Maintenance	N_{multi}
1	6	ϕ (σ)	0.24 (0.05)	6	0.06 (0.10)	1
		A (σ)	0.56 (0.22)		0.94 (0.63)	
2	6	ϕ (σ)	0.25 (0.12)	4	-0.03 (0.07)	1
		A (σ)	0.35 (0.16)		1.03 (0.45)	
3	6	ϕ (σ)	0.23 (0.11)	6	0.18 (0.11)	5
		A (σ)	1.36 (0.22)		0.70 (0.28)	

The results presented in Table C3 showed that neither providing feedback nor providing perceptual anchors was helpful in facilitating multidimensional categorization in the maintenance phase. The ϕ dropped to 0 in both conditions and the number of listeners categorizing multidimensionally dropped to 1. Providing a verbal description of the category structures and reminding the listeners of the importance of using both dimensions, however, was helpful as both the mean ϕ and the number of listeners using both dimensions in the maintenance phase of Condition 3 shows.

These observations were confirmed by the statistical evaluation of the ϕ's of the learning and maintenance phases of the conditions. In the learning phase of Conditions 1 and 2, ϕ differs significantly from 0 and $\frac{1}{2}\pi$ ($t_{min} = 5.2$, $p < 0.05$). In the

maintenance phase of these conditions, however, ϕ only differs significantly from $\frac{1}{2}\pi$ ($t_{min} = 11.2$, $p < 0.05$) showing a return to unidimensional categorization. Condition 3 has similar results in the learning phase with ϕ differing from both 0 ($t[5] = 4{,}8$, $p < 0.05$) and $\frac{1}{2}\pi$ ($t[5] = -5{,}7$, $p < 0.05$). However, the maintenance phase of Condition 3 shows multidimensional categorization also: ϕ differs significantly from both 0 ($t[5] = 4.0$, $p < 0.05$) and $\frac{1}{2}\pi$ ($t[5] = -7.4$, $p < 0.05$).

The consistency measure A differed significantly from zero in all phases in all conditions ($t_{min} = 3.7$, $p < 0.05$) and was used to directly evaluate the differences between the conditions. The ANOVA showed a statistical difference between the conditions in the learning phase ($F[2,15] = 40.98$, $p < 0.05$), but not in the maintenance phase ($F[2,15] = 0.78$, n.s.). Post hoc testing (Tukey HSD) confirmed that Condition 1 and 2 form a homogeneous subset opposed to Condition 3 in the learning phase but not in the maintenance phase. The effect of the written explanation is thus present in the comparison of the ϕ's, but not in the consistency measure A.

Discussion

The manipulations of Experiment 2 were aimed at improving categorization performance in the maintenance phase by changing the conditions in the learning phases. In Condition 1, listeners received right/wrong feedback on every trial to help them learn to integrate the two relevant dimensions. Condition 2 added perceptual anchors to the feedback to further help category learning. Finally, Condition 3 added a written instruction describing the category structures and stressing the importance

of the use of both dimensions. Only with the last manipulation were listeners able to maintain the categorization rule they successfully applied in all learning phases. Since verbal instruction about the separation of phonetic categories can hardly be thought of as an ecologically valid approach, this procedure was not used in the category learning experiments of Chapter 2, 3, and 4. The aim of the research presented there was to discover what listeners were able to learn when left to their own devices, whereas Condition 3 aimed at finding the upper limits of performance in the maintenance phase.

Conclusions

As has been shown in this appendix and throughout the thesis, listeners *are* sensitive to the distributional properties of the stimuli. The categorization rules they apply in the learning phases and the change in their categorizations in the maintenance phase indicate as much. However, the differences in range and standard deviation created in Experiment 1 did not succeed helping listeners maintain the categorization rule they applied in the learning phase, neither did adding right/wrong feedback or perceptual anchors. The only manipulation that succeeded in helping listeners maintain their acquired categorization tendencies was providing them with a clear instruction to use both dimensions and literally pointing the category structure out to them.

Appendix D

Incidental category learning[18]

18 The experiment presented in this appendix was carried out with financial support from the Dutch Scientific Council. We further thank Keith Kluender, University of Wisconsin, Madison for financial and other assistance with these experiments.

Introduction

Learning without the aid of trial-by-trial feedback (unsupervised learning) can take many forms. In the experiments that were presented in this thesis, the difference between supervised and unsupervised learning was usually the absence of feedback on the categorization of the listener. However, other forms of unsupervised learning are also possible: observational learning with or without cues (as was presented in Appendix A) and *implicit* learning; learning that takes place without the listener being (explicitly) aware of it. This last form of learning could be one of the ways in which speech categories are acquired (since infants in all likelihood lack an explicit reasoning system) and it is the subject of this appendix.

This experiment presented listeners with three categories instead of two. The listeners had to explicitly differentiate (in an oddball task) one category from the other two, while the other two categories were never explicitly contrasted with one another. However, listeners *did* observe both of them in contrast to the third category. This procedure was meant to create a multidimensional categorization tendency in listeners, possibly combined with categorical perception effects.

Method

Subjects

Thirty-six students from the University of Wisconsin, Madison, took part in the experiment and were given course credit in return. All of them signed a consent form at the beginning of the experiment and none of them reported any hearing problems.

Stimuli

Three versions of the experiment were run, each with slightly different category structures in terms of the range and standard deviations (see Appendix C). The stimuli were the inharmonic sound complexes differing in duration in formant frequency introduced in Chapter 2. However, this time there were three instead of two categories. Figure D1 shows the three categories used in the learning phase in their multidimensional perceptual space (Category A being the one in the lower left corner, category B being the middle one, and category C being the one in the upper right corner).

The difference between the means of each category was set to 20 just noticeable differences. Each category contained 112 stimuli which brought the total number of stimuli in the learning phase to 336.

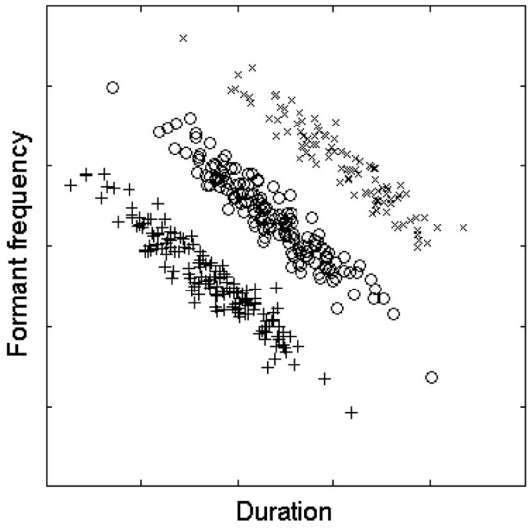

Figure D1. Three multidimensional category structures in a twodimensional space.

In the discrimination phase, listeners were presented with 14 stimuli drawn from an equidistantly spaced diagonal line from the mean of category A to the mean of category B. The maintenance stimuli were constructed using the equidistantly spaced grid introduced in Chapter two. The stimulus characteristics of the maintenance grid as well as those of category A and B were identical to the stimuli used in Chapter two. Category C differed in this respect, because it was located in a different (higher) area of perceptual space.

Procedure

The experiment consisted of a learning phase, a discrimination phase , and a classification phase. In the learning phase, listeners heard four sounds: three from either category A or B, and one from C (the category in the upper right corner of Figure D1). One of the four sounds was the "odd one out" and listeners were expected to identify which one by pressing one of four buttons. After their response,

a light above the button indicated which sound was the odd one out. There were 200 trials in the learning phase, 100 with three exemplars of category A and one of category C (in random combinations), and another 100 with three exemplars of category B and one of category C (in random combinations). The training phase took about 15 minutes to complete.

After the training phase, listeners entered a discrimination phase in which they had to discriminate stimuli ranging between the means of categories A and B in an AXB paradigm. In this procedure, listeners were presented with three sounds and either the first or the last two were the same (aab, abb, bba, or baa). Listeners were asked to indicate whether the first or the last pair was the same by pressing a button. No feedback was given on their responses. The discriminations were made with a stepsize of three, meaning that it was tested whether listeners could discriminate between stimulus one and stimulus four, between two and five, et cetera. Each of the 44 possible triplets was presented four times, resulting in 176 stimuli. The discrimination task took about 12 minutes to complete.

Finally, after the discrimination phase, listeners entered the maintenance phase where they categorized stimuli drawn from the equidistantly spaced grid (positioned between the means of the categories A and B). Each of the 49 stimuli was presented four times, resulting in 196 stimuli. The maintenance phase took about ten minutes to complete.

Results and discussion

Learning phase and discrimination phase

Three versions of the experiment were run, each with a slightly different value for the just noticeable difference and the distance between the means of the categories (see Appendix C). The results for the training, discrimination, and maintenance phase of the three versions were similar, so the data were pooled. Figure D2 and Figure D3 show the results for the learning and discrimination phase respectively.

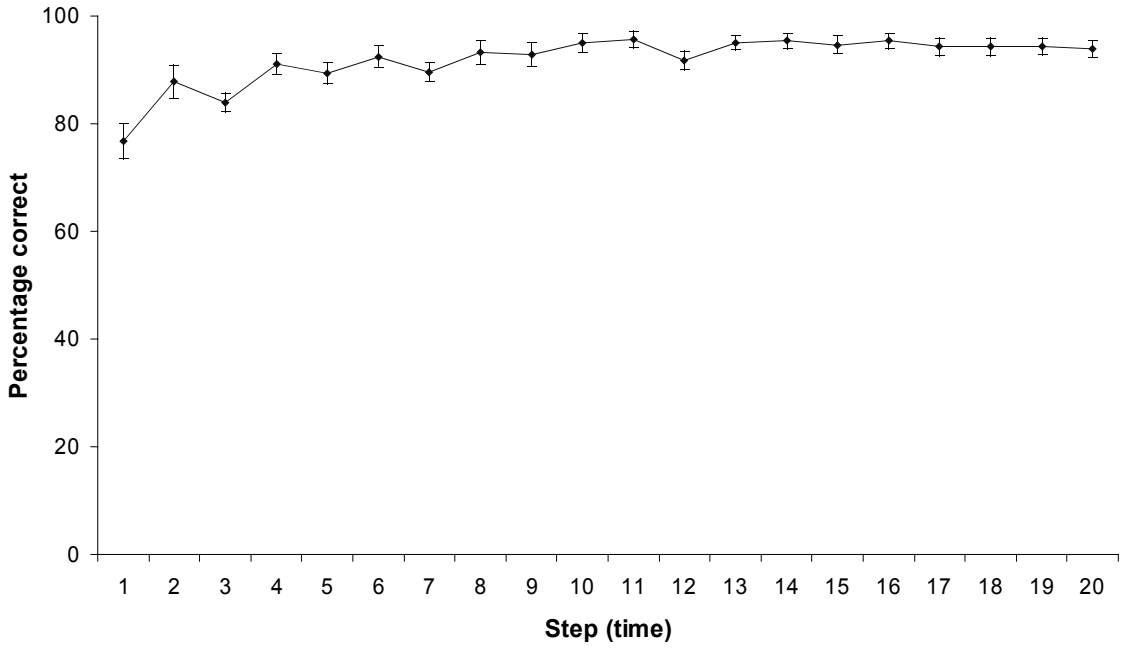

Figure D2. Percentage correct over time in the learning phase.

As Figure D2 shows, subjects were quite good at determining which stimulus was the odd one out. Percent correct starts around 70 and reaches 90 at the end of the learning phase. Figure D3 shows the discrimination results. Although

discrimination is certainly above chance level at every step (t_{min} [27] = 3.2, $p < 0.05$), the peak in discriminatory ability in between the two categories that is thought to be a property of categorical perception, is absent.

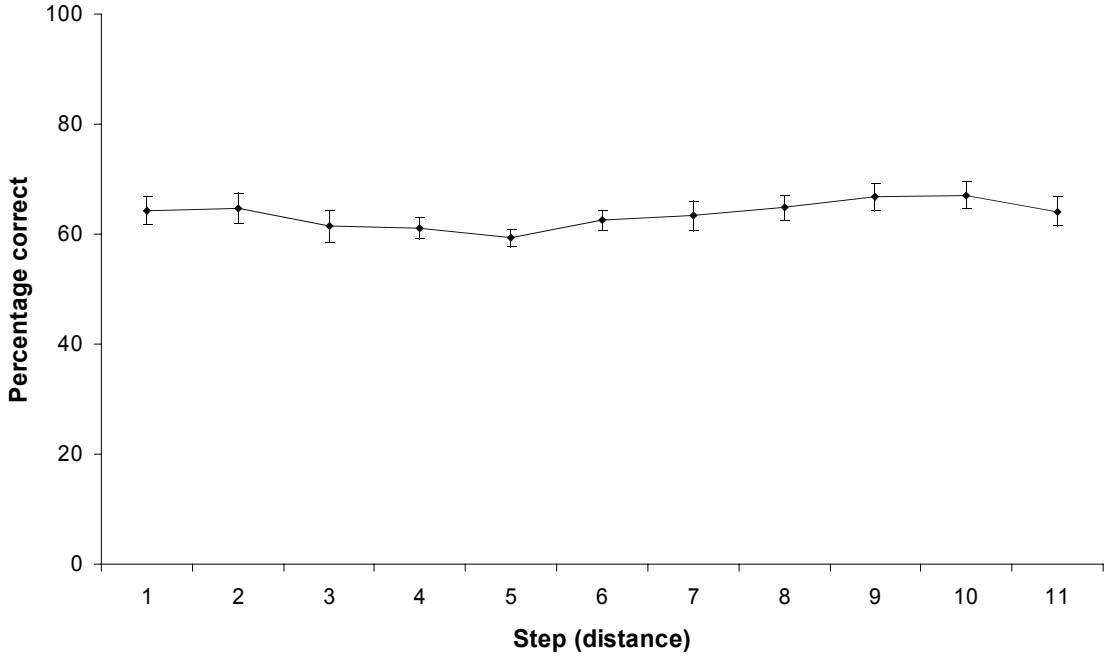

Figure D3. Percentage correct in the discrimination phase.

Maintenance phase

The polar transformations of the logistic regression weights for duration and frequency as well as the number of participants using both dimensions in their categorization are presented in table D1.

Table D1.
Logistic regression results of the maintenance phase.

N	A	ϕ	N_{multi}
28	1.08 (0.76)	0.14 (0,40)	5

The consistency measure A was high and significantly different from zero ($t[27] = 7.5$, $p < 0.05$) but the mean angle ϕ differed only marginally significant from 0 ($t[27] = 1.9$, $p < 0.06$) and highly significantly from $\frac{1}{2}\pi$ ($t[27] = -4.74$, $p < 0.000$). Together with the small number of listeners using both dimensions significantly, this indicates that the often observed preference for duration was not altered by the experimental manipulations.

Taken together, the results of the experiment suggest that the three category oddball task is a doable one, but does not result in multidimensional categorization. Either multidimensional category learning does not happen implicitly or the task does not tap into implicit learning mechanisms enough. We decided on the latter and did not use this procedure in subsequent experiments.

Samenvatting in het Nederlands

Hoe baby's de klanken van hun moedertaal leren herkennen is onderzoekers nog steeds een raadsel. Dit proefschrift probeerde enkele van de processen betrokken bij het leren van klanken van zowel een eerste als een tweede taal in kaart te brengen.

Het complexe probleem waar zowel baby's als volwassen tweede taalverwervers mee geconfronteerd worden werd hier teruggebracht tot het kunnen onderscheiden van twee categorieën. Waar spraakklanken (fonetische categorieën) op allerlei manieren van elkaar verschillen, verschilden de klanken in dit onderzoek steeds in een of twee eigenschappen (dimensies) van elkaar. De variatie in beide dimensies werd gemanipuleerd, zodat ofwel één of beide dimensies relevant waren. Dit valt goed te zien in Figuur 2.1.

Omdat baby's de klanken noodzakelijkerwijs *zonder* feedback moeten leren (ze gaan pas klanken produceren nadat ze al in staat zijn deze te herkennen), werd de rol van feedback bij het leren van deze klanken onderzocht door experimenten mét feedback te vergelijken met experimenten zónder feedback.

Tenslotte werd onderzocht wat de invloed is van de samenstelling van het klankpatroon van de moedertaal op het leren van de klanken van een tweede taal door volwassenen.

Alle experimenten werden gedaan met volwassen deelnemers. Om het leren van de eerste klanken zoals dat bij baby's plaatsvindt, te onderzoeken werd gebruik

gemaakt van klanken die niet als spraak werden waargenomen, maar daar wel erg op leken. Om het leren van de klanken van een tweede taal te onderzoeken, werd gebruik gemaakt van Spaanstalige en Engelstalige luisteraars die klanken uit het Nederlands moesten leren categoriseren.

Zowel bij de niet-spraakklanken als bij de spraakklanken waren de dimensies die gebruikt werden voor het maken van de verschillende categorieën de *duur* van het geluid (in milliseconden, gemiddeld 150 milliseconden) en de piekfrequentie van de eerste formant (de *resonantiepiek*).

Het leren van niet-spraakklanken

De experimenten in hoofdstuk 2 onderzochten het leren van niet-spraak categorieën met behulp van supervisie, waarbij de deelnemers feedback kregen op elke reactie (goed of fout). In experiment 1 werden luisteraars getraind om klanken te categoriseren die zowel een relevante als een irrelevante dimensie van variatie hadden (zie de eerste twee bovenpanelen van Figuur 2.1). De resultaten lieten zien dat deze categoriestructuren eenvoudig te leren waren met behulp van feedback. Het *vasthouden* van het geleerde in een testfase zonder feedback en zonder distributionele informatie (zie het derde bovenpaneel van Figuur 2.1) was een stuk moeilijker. Vooral de dimensie resonantiepiek werd door de luisteraars nauwelijks nog gebruikt in de testfase.

Experiment 2 onderzocht het leren van een multidimensionele structuur, waarin zowel duur als resonantiepiek belangrijk zijn voor het onderscheid tussen de categorieën (zie de eerste twee onderpanelen van Figuur 2.1), weer met behulp van feedback is onderzocht in experiment 2 van hoofdstuk 2. Een dergelijk

multidimensioneel onderscheid was veel moeilijker te leren dan wanneer er maar één dimensie relevant was, zelfs wanneer luisteraars op iedere keuze die ze maakten feedback kregen. Hoewel bijna alle luisteraars beide dimensies uiteindelijk leerden te gebruiken in de trainingsfase, bleek het geleerde in de testfase wederom fragiel. Net als in experiment 1 gebruikten luisteraars in de testfase het liefst één dimensie, waarbij zij de voorkeur gaven aan duur boven resonantiepiek.

De testfase verschilde van de trainingsfase op twee manieren: er was geen feedback meer én de distributionele informatie die aanwezig was in de trainingsstimuli was afwezig in de testfase (zie voor een illustratie Figuur 2.1). Experiment 3 onderzocht welk van deze twee bronnen van informatie verantwoordelijk was voor het (on)vermogen van de luisteraars om het geleerde multidimensionele onderscheid te handhaven in de testfase. De stimuli in de testfase van experiment 3 waren namelijk identiek aan de stimuli uit de trainingsfase en bezaten dus nog wel distributionele informatie. De testfase verschilde nu alleen nog maar van de trainingsfase in de afwezigheid van feedback. In deze testfase mèt distributionele informatie slaagden luisteraars er wel in om volgens het geleerde onderscheid te categoriseren. Uit dit resultaat bleek dat luisteraars erg gevoelig waren voor de aan- en afwezigheid van distributionele informatie en hun categorisatie daar vrijwel meteen op aanpasten.

In hoofdstuk 3 werd het leren van dezelfde categoriestructuren en klanken als in hoofdstuk 2 onderzocht, maar nu zónder feedback.

In experiment 1 was slechts één van de beide dimensies relevant, terwijl de variatie in de andere dimensie niet van belang was (voor categorie lidmaatschap). Verrassend was dat luisteraars in staat waren om zonder hulp van feedback te ontdekken welke dimensie relevant was en om deze dimensie vervolgens te

gebruiken in hun categorisatie. In de trainingsfase bleek het ontdekken van piekfrequentie als relevante dimensie iets makkelijker dan duur, terwijl de luisteraars het in de testfase juist lastig vonden om duur te negeren en de variatie in resonantiepiek te gebruiken. Net als in hoofdstuk 2 was er in de testfase dus een voorkeur voor de dimensie duur.

In experiment 2 werd onderzocht of luisteraars ook multidimensionele stimuli konden leren categoriseren zonder hulp van feedback. De prestaties van individuele proefpersonen verschilden sterk, maar toch was een aanzienlijk deel van de luisteraars gevoelig voor de distributionele eigenschappen van de geluiden en maakten ze dus in hun categorisatie gebruik van beide dimensies. In vergelijking met Experiment 1, waar één dimensie relevant was en de ander genegeerd moest worden, was het leereffect echter veel kleiner.

Toch blijkt uit de vergelijking van de experimenten uit hoofdstuk 2 en hoofdstuk 3 dat de verschillen tussen leren met en leren zonder feedback eerder kwantitatief dan kwalitatief zijn. Leren zonder supervisie gaat langzamer en moeizamer, maar verschilt verder niet van leren met feedback: categoriestructuren waarin beide dimensies relevant zijn worden in beide gevallen moeilijker gevonden dan structuren met een relevante en een irrelevante dimensie.

Het leren van spraakklanken

In de experimenten in hoofdstuk 4 werden andere categorieën geleerd dan in de eerdere hoofdstukken: in plaats van niet spraak worden de luisteraars hier blootgesteld aan door een computer gegenereerde Nederlandse klinkers. Deze klinkers, de *eu*, *uu* en *u* (in fonetisch schrift de /ø/, /y/, en /Y/), worden voornamelijk

van elkaar onderscheiden door dezelfde dimensies als in de vorige hoofdstukken: duur van de klinker en eerste resonantiepiek. De *eu* (feut) is langer dan de *u* (fut), maar verder hetzelfde, de *uu* van fuut verschilt vooral van de *eu* van feut in piekfrequentie en voor het onderscheid van de *eu* en de *u* zijn beide dimensies noodzakelijk. De luisteraars waren ofwel Spaanstalig (Experiment 1, 2 en 3) ofwel Engelssprekenden uit de VS (Experiment 3 en 4). In beide talen zijn deze klinkers onbekend. Het onderscheid tussen de categorieën werd zowel met als zonder feedback geleerd.

Experiment 1 liet zien dat Spaanstalige luisteraars klanken die van elkaar verschillen op één relevante dimensie met behulp van feedback konden leren. Dat konden ze voor beide dimensies, al hadden ze een voorkeur voor piekfrequentie. Deze voorkeur was het duidelijkst in de testfase.

In experiment 2 moesten de Spaanse luisteraars hetzelfde onderscheid zonder feedback leren. Door de afwezigheid van feedback konden ze alleen gebruik maken van de distributionele eigenschappen van de stimuli. Nu waren de luisteraars niet in staat het verschil tussen beide categorieën te leren. Ook hier was er weer duidelijk een voorkeur om piekfrequentie te gebruiken, of die dimensie nou relevant was of niet. Wanneer piekfrequentie relevant was in de trainingsfase gebruikten luisteraars deze dimensie zeer sterk in de testfase, maar wanneer duur relevant was in de trainingsfase gebruikten ze deze dimensie nauwelijks in de testfase. Het is mogelijk dat de fonologische eigenschappen van het Spaans hier een rol in spelen: piekfrequentie is daar erg belangrijk, terwijl er in het Spaans geen klinkers zijn die onderscheiden worden met behulp van duur.

Als de fonologie van de moedertaal verantwoordelijk is voor de moeite die de Spaanse luisteraars hebben met het gebruik van duur in hun categorisatie, dan

zouden sprekers van een taal waar duur wél een belangrijke dimensie is daar minder moeite mee hebben. In experiment 3 werd daarom onderzocht hoe sprekers van het Amerikaans-Engels het duur onderscheid tussen *eu* (feut) en *u* (fut) leerden (mét feedback, net als de Spaanstaligen in experiment 1). Het bleek dat de Engelse luisteraars dit op duur gebaseerde onderscheid veel beter konden leren dan de Spaande luisteraars uit experiment 1 en 2. Dit resultaat ondersteunt de hypothese dat de fonologie van de moedertaal een belangrijke rol speelt bij het leren van nieuwe klanken.

Tot slot onderzocht experiment 4 het leren van het multidimensionele onderscheid tussen *uu* en *u* met feedback. De resultaten lieten zien dat sprekers van het Amerikaans-Engels (een taal waarin beide dimensies een belangrijke rol in de fonologie hebben) moeite hadden om beide dimensies te leren gebruiken, zélfs met de hulp van feedback. Toch bleek uit de resultaten ook dat deze luisteraars wel gevoelig waren voor de beide bronnen van informatie (feedback en distributionele eigenschappen), want aan het einde van de trainingsfase maakte de helft van de luisteraars gebruik van beide dimensies bij het categoriseren. In de testfase bleek dit leren wel weer fragiel en vielen de luisteraars terug op het gebruik van één dimensie (meestal piekfrequentie) in hun categorisatie.

Conclusies

Samenvattend leiden de resultaten beschreven in dit proefschrift tot een drietal hoofdconclusies.

Ten eerste zijn de verschillen in het leren van klanken met en zonder feedback kwantitatief en niet kwalitatief van aard. Ten tweede zijn luisteraars gevoelig van

voor distributionele informatie bij het leren van klanken, zelf wanneer zij leren zonder feedback. Ten derde speelt de fonologie van de moedertaal een belangrijke rol bij het leren van de klanken van een tweede taal.

Curriculum vitae

Martijn Goudbeek werd geboren in Enschede op 31 December 1975 . Hij doorliep de OBS 't Vastert van 1980 tot 1988 en de haalde in 1994 een vwo diploma aan de Scholengemeenschap Zuid (thans het Stedelijk Lyceum) te Enschede. In 1994 begon hij met de opleiding psychologie aan de Katholieke Universiteit Nijmegen (thans Radboud Universiteit Nijmegen). Na een klinische stage bij het Wilhelmina Kinderziekenhuis rondde hij in 1999 deze opleiding af bij de sectie Neuro- en Revalidatiepsychologie met een afstudeerscriptie over elektrische huidweerstand en volgehouden aandacht bij ADHD kinderen. Een jaar later sloot hij zijn opleiding tot wijsgeer van een wetenschapsgebied aan de Faculteit der Wijsbegeerte en Theologie af met een studie naar computationale en dynamische opvattingen over mentale representaties. In zijn studietijd was hij lid van de opleidingscommissie wijsbegeerte en van de studievereniging Neuro- en Revalidatiepsychologie “Homunculus”. Daarnaast vervulde hij diverse student-assistentschappen. Na een half jaar ADHD onderzoek op Trinity College, Dublin begon hij zijn promotietraject op het Max Planck Instituut voor Psycholinguïstiek (2001-2006). In het kader van een NWO reisbeurs verbleef hij zes maanden aan de Universiteit van Wisconsin, Madison om daar een deel van het promotieonderzoek uit te voeren. Momenteel werkt hij in Genève als postdoctoraal onderzoeker aan de Geneva Emotion Research Group.

MPI series in psycholinguistics

1. The electrophysiology of speaking: Investigations on the time course of semantic, syntactic, and phonological processing. Miranda van Turennout
2. The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsaggital articulography. Niels O. Schiller
3. Lexical access in the production of ellipsis and pronouns. Bernadette M. Schmitt
4. The open-/closed-class distinction in spoken-word recognition. Alette Haveman
5. The acquisition of phonetic categories in young infants: A self-organising artificial neural network approach. Kay Behnke
6. Gesture and speech production. Jan-Peter de Ruiter
7. Comparative intonational phonology: English and German. Esther Grabe.
8. Finiteness in adult and child German. Ingeborg Lasser
9. Language input for word discovery. Joost van de Weijer
10. Inherent complement verbs revisited: Towards an understanding of argument structure in Ewe. James Essegbey
11. Producing past and plural inflections. Dirk Janssen
12. Valence and transitivity in Saliba: An Oceanic language of Papua New Guinea. Anna Margetts
13. From speech to words. Arie van der Lugt

14. Simple and complex verbs in Jaminjung: A study of event categorisation in an Australian language. Eva Schultze-Berndt
15. Interpreting indefinites: An experimental study of children's language comprehension. Irene Krämer
16. Language-specific listening: The case of phonetic sequences. Andrea Weber
17. Moving eyes and naming objects. Femke van der Meulen
18. Analogy in morphology: The selection of linking elements in Dutch compounds. Andrea Krott
19. Morphology in speech comprehension. Kerstin Mauth
20. Morphological families in the mental lexicon. Nivja H. de Jong
21. Fixed expressions and the production of idioms. Simone A. Sprenger
22. The grammatical coding of postural semantics in Goemai (a West Chadic language of Nigeria). Birgit Hellwig
23. Paradigmatic structures in morphological processing: Computational and cross-linguistic experimental studies. Fermín Moscoso del Prado Martín
24. Contextual influences on spoken-word processing: An electrophysiological approach. Daniëlle van den Brink
25. Perceptual relevance of prevoicing in Dutch. Petra M. van Alphen
26. Syllables in speech production: Effects of syllable preparation and syllable frequency. Joana Cholin
27. Producing complex spoken numerals for time and space. Marjolein Meeuwissen
28. Morphology in auditory lexical processing: Sensitivity to fine phonetic detail and insensitivity to suffix reduction. Rachèl J. J. K. Kemps
29. At the same time...: The expression of simultaneity in learner varieties. Barbara Schmiedtová

30. A grammar of Jalonke argument structure. Friederike Lüpke
31. Agrammatic comprehension: An electrophysiological approach. Marlies Wassenaar
32. The structure and use of shape-based noun classes in Miraña (North West Amazon). Frank Seifart
33. Prosodically-conditioned detail in the recognition of spoken words. Anne Pier Salverda
34. Phonetic and lexical processing in a second language. Mirjam Broersma
35. Retrieving semantic and syntactic word properties: ERP studies on the time course in language comprehension. Oliver Müller
36. Lexically-guided perceptual learning in speech processing. Frank Eisner
37. Sensitivity to detailed acoustic information in word recognition. Keren B Shatzman
38. The Relationship between spoken word production and comprehension. Rebecca Özdemir
39. Disfluency: Interrupting speech and gesture. Mandana Seyfeddinipur
40. The acquisition of phonological structure: Distinguishing contrastive from non-contrastive information. Christiane Dietrich
41. Cognitive cladistics and the relativity of spatial cognition. Daniel Haun
42. Acquiring auditory categories. Martijn Goudbeek