Journal of Experimental Psychology:
Human Perception and Performance
2006, Vol. 32, No. 3, 733–754

0096-1523/06/$12.00 DOI: 10.1037/0096-1523.32.3.733

Categorization of Sounds

Roel Smits
Max Planck Institute for Psycholinguistics

Joan Sereno and Allard Jongman
University of Kansas

The authors conducted 4 experiments to test the decision-bound, prototype, and distribution theories for the categorization of sounds. They used as stimuli sounds varying in either resonance frequency or duration. They created different experimental conditions by varying the variance and overlap of 2 stimulus distributions used in a training phase and varying the size of the stimulus continuum used in the subsequent test phase. When resonance frequency was the stimulus dimension, the pattern of categorization-function slopes was in accordance with the decision-bound theory. When duration was the stimulus dimension, however, the slope pattern gave partial support for the decision-bound and distribution theories. The authors introduce a new categorization model combining aspects of decision-bound and distribution theories that gives a superior account of the slope patterns across the 2 stimulus dimensions.

Keywords: response, auditory, categorization, models, noisy encoding noisy activation (NENA)

The categorization of sounds plays an important role in everyday life. On a daily basis, people categorize sounds in their environment as belonging to such basic events as a telephone ringing, a baby crying, or the doors of a train slamming shut. The correct categorization of sounds may even be of vital importance, such as the honking of an approaching car. Finally, the recognition of the sounds and words of spoken language is a form of auditory categorization that occupies a significant portion of most people's waking hours.

Despite its importance, auditory categorization has received relatively little attention in the psychological literature. An overwhelming majority of studies on perceptual categorization have been devoted to the categorization of simple visual stimuli, such as line segments of variable lengths and orientations. Three distinct theories of visual categorization have featured most prominently in the recent literature. Prototype theory (Rosch, 1973; Smith & Minda, 2000) assumes that stimuli are categorized on the basis of their similarity to category prototypes stored in memory. A category prototype is generally defined as the average, or most typical, member of a category. Exemplar theory (Nosofsky, 1986), conversely, denies the explicit use of category prototypes. In its extreme formulation, exemplar theory assumes that categorization is based on a comparison of the stimulus with all previously categorized exemplars of all categories. Finally, decision-bound theory (Ashby & Perrin, 1988) assumes that categorization is based on the comparison of the perceptual effect of a stimulus with category boundaries stored in memory. Explicit connections between the fields of speech perception and visual categorization have been sparse. Nevertheless, the hypotheses for the categorization of phonemes that have been proposed over the years are similar to those for visual categorization, albeit more qualitative and mathematically less well developed.

One of the most popular research methodologies in speech perception is the phoneme categorization experiment. In phoneme categorization experiments, listeners are presented with naturally produced or synthetic speech sounds and are asked to assign the sounds to phonetic categories. In the most widely used version of the phoneme categorization experiment, synthetic stimuli are used in which one or more acoustical parameters of interest, such as formant frequencies or silence durations, are systematically varied in a number of discrete steps of equal size to form a stimulus continuum.

Researchers from Haskins Laboratories (New Haven, CT) pioneered the use of stimulus continua for investigating speech perception, showing, among other things, that the perceived phoneme is generally influenced by many acoustic parameters distributed over a wide temporal window (Cooper, Delattre, Liberman, Borst, & Gerstman, 1952). Since these landmark studies, researchers have used stimulus continua to investigate many basic aspects of speech perception, such as the unit of recognition (Nearey, 1997), dependencies in the categorization of successive phonemes (Massaro & Cohen, 1983; Smits, 2001), the influence of speaking rate on phonetic categorization (Volaitis & Miller, 1992), and the relative weights of various acoustic cues to particular phonetic distinctions (Ainsworth, 1968).

Despite the popularity of the phoneme categorization paradigm, investigators still do not fully understand how phonetic categories are represented and what listeners actually do in phoneme categorization experiments. How do people categorize speech sounds? In the present study, we take a first step in answering this question, focusing on the categorization of synthetic nonspeech sounds.

Roel Smits, Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands; Joan Sereno and Allard Jongman, Department of Linguistics, University of Kansas.

We are grateful to Aoju Chen, Tau van Dijk, Manon van Laer, Elske Hissink, Jessica Pas, and Margret van Beuningen for running the experiments and to Jim Miller and Terry Nearey for useful comments.

Correspondence concerning this article should be addressed to Roel Smits, Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, the Netherlands. E-mail: roel.smits@mpi.nl

In the past, researchers have proposed several hypotheses concerning the representation and categorization of speech sounds. Fueled by the "categorical perception" controversy, investigators mainly analyzed and discussed early speech perception experiments in terms of the boundaries between phonetic categories (e.g., Liberman, Harris, Hoffman, & Griffith, 1957). They suggested that what listeners do in phonetic categorization experiments is evaluate on which side of the relevant phonetic boundary the perceptual effect of the incoming stimulus is located. One can view this hypothesis as a phonetic implementation of decision-bound theory. Experimenters have also hypothesized that phoneme categories may be represented by prototypes (e.g., Kuhl, 1991; Oden & Massaro, 1978). In the context of speech perception, prototype theory assumes that listeners in phonetic categorization experiments compute the similarity of an incoming stimulus to each of the relevant category prototypes and categorize the stimulus on the basis of these similarities.

More recently, Nearey and colleagues and Miller (1994) supported a more elaborate view of phonetic category representation that contains more information than just that found for prototypes. Miller claimed that representations of phonetic categories are essentially graded. This claim is based on the finding that category members vary in their perceived category goodness. Miller's position must be interpreted as mainly contrasting with the classical categorical perception concept, whereby members of the same category are thought to be perceptually entirely equivalent. In principle, graded category structure may derive from a basic prototype representation, as Miller acknowledged, whereby members close to the prototype are judged better exemplars than members further away from the prototype. Miller did suggest, however, that phonetic categories actually incorporate distributional information. Nearey and colleagues (Andruski & Nearey, 1992; Assman, Nearey, & Hogan, 1982; Hillenbrand & Nearey, 1999; Neary & Assman, 1986) took a more quantitative stance. Using the normal a posteriori probability model, they modeled listeners' representations of vowel categories as multidimensional Gaussian distributions and showed that a posteriori vowel probabilities based on their model gave very good predictions of listeners' categorization for a given set of vowel stimuli. Henceforth, we indicate the class of theories that assume that phonetic categories are represented as distributions (of some form) as the *distribution theory*.

Finally, a contemporary hypothesis concerning phonetic categorization simply denies the existence or involvement of a sublexical layer in human speech recognition, instead assuming that words are represented by many, possibly all, previously encountered exemplars of the word (Goldinger, 1997; Johnson, 1997a, 1997b). Henceforth, we indicate this categorization theory as the *exemplar theory*. From this perspective, phonemes are not explicitly represented at all, and exactly what listeners do in phonetic categorization tasks is not of central importance in understanding how human speech recognition works. Although it is logically possible to formulate an exemplar theory of speech perception with a sublexical layer containing phoneme exemplars, such a theory, to our knowledge, has not been proposed.

Despite 20 years of experimenting, the issue of the basic mechanisms underlying perceptual categorization has not been resolved. In speech perception research, the number of experiments addressing the basic mechanisms in phonetic categorization is small compared with research on visual categorization (see, e.g., Maddox & Ashby, 1998; Nosofsky, 1998). Although there is currently no consensus in the fields of speech perception and general perceptual categorization, theories of general perceptual categorization have the advantage that they are mathematically fully developed and are the subject of intense experimental evaluation. In the present research, we therefore attempt to apply some of the methods and models of the visual categorization literature to the problem of phonetic categorization.

A major problem hindering a direct transfer of the methods to the speech domain is that experimenters do not have any control over the "training corpus" to which participants have been exposed. Throughout their life, while hearing other people speak, adult listeners have heard many instances of the phonemes in their language. Both the basic categorization mechanism used by listeners and a number of system "parameters" (boundary locations, prototype locations, or distribution covariance matrices) may be based on this—unknown—training corpus. This poses two problems. First, the training corpus differs among listeners, and, second, we cannot freely manipulate the training corpus to test certain theoretical predictions. One methodological approach capable of overcoming the difficulties stemming from the lack of control over the training corpus is to study the categorization of nonspeech sounds. In the present study we use this methodology.

Our experimental approach uses the method of externally distributed stimuli (EDS). We build on the idea of Lee and Zentall (1966) that variation of the training distributions should cause different categorization theories to predict different patterns of categorization-function slopes. However, we wanted to create an experimental situation that was similar to that of the phonetic categorization experiment. This led us to adopt a methodology with quite distinct training and test phases. In the experiments we report, we used the EDS method for training the participants, who received feedback after every trial. In the test phase, conversely, we used a stimulus continuum and gave no feedback, thus mimicking the standard phonetic categorization task.

The stimuli used in the present study differ in a number of regards from natural speech sounds. First, they are nonspeech sounds that resemble speech sounds in certain crucial aspects. We believe that the use of nonspeech is warranted because it is the best way to control for differences in participants' previous exposure to speech. Second, our stimuli differ on only a single dimension, whereas natural speech varies along many dimensions (including frequency, duration, and amplitude). Although we acknowledge these differences, we adopted the present strategy of studying simple auditory stimuli because, relative to visual categorization, phonemic categorization is, as yet, not well understood. Only when an account of the categorization of relatively simple auditory stimuli has been developed can research begin to address the categorization of more complex signals that will increasingly resemble speech sounds. The current study should therefore be considered a first step toward understanding the process of phonetic categorization.

We organize the remainder of the article as follows. First, we present the general methodology for all four reported experiments. Next, we define the categorization theories mathematically and derive their predictions of listeners' categorization functions. Subsequently, we present four experiments that test the theories. We then describe quantitative model-based analyses, including a new categorization model. Finally, we discuss the results and interpret them within the context of speech perception.

General Method

In all the experiments we report, we trained listeners on a categorization problem involving two categories, A and B. Categories A and B were defined by overlapping one-dimensional Gaussian probability density functions (pdfs), pdf_A and pdf_B, characterized by means μ_A and μ_B and standard deviations σ_A and σ_B. On a given trial in the training phase, we randomly drew a stimulus from pdf_A or pdf_B and played it to the listener. He or she had to label the stimulus as either A or B, after which we gave visual feedback on the correct response. After completing the training phase, listeners entered the test phase. In this phase, they performed the same task, but this time without getting feedback.

As we indicated in the introduction, the methodology that we used for distinguishing among the four models of categorization had two essential features. First, according to the EDS technique of Lee and Zentall (1966), we used four different training conditions. We created these by orthogonally combining two levels of variance and two levels of overlap of pdf_A and pdf_B. The leftmost column of Figure 1 gives a graphical representation of the four training conditions. Within conditions, σ_A and σ_B were always equal. In Conditions 1, 2, 3, and 4, the distance $\Delta\mu$ between means μ_A and μ_B was set to 5, 10, 10, and 20 just-noticeable differences (jnds), respectively, on the associated psychological dimension. Standard deviations σ_A and σ_B were 3.704 jnds in Conditions 1 and 2 and 7.407 jnds in Conditions 3 and 4. As a result, the overlap of pdf_A and pdf_B was large in Conditions 1 and 3 and small in Conditions 2 and 4, with theoretically optimal classification rates (according to a noise-free boundary-based classification rule) of 75.5% in Conditions 1 and 3 and 91.8% in Conditions 2 and 4. Table 1 summarizes the numerical parameters of the four training conditions.

pdf_A and pdf_B were represented by 110 stimuli each. As did Lee and Zentall (1966), we sampled parameter values in such a way that the interval between any pair of consecutive parameter values corresponded to a constant probability interval on the cumulative distribution function associated with the pdf of the category.

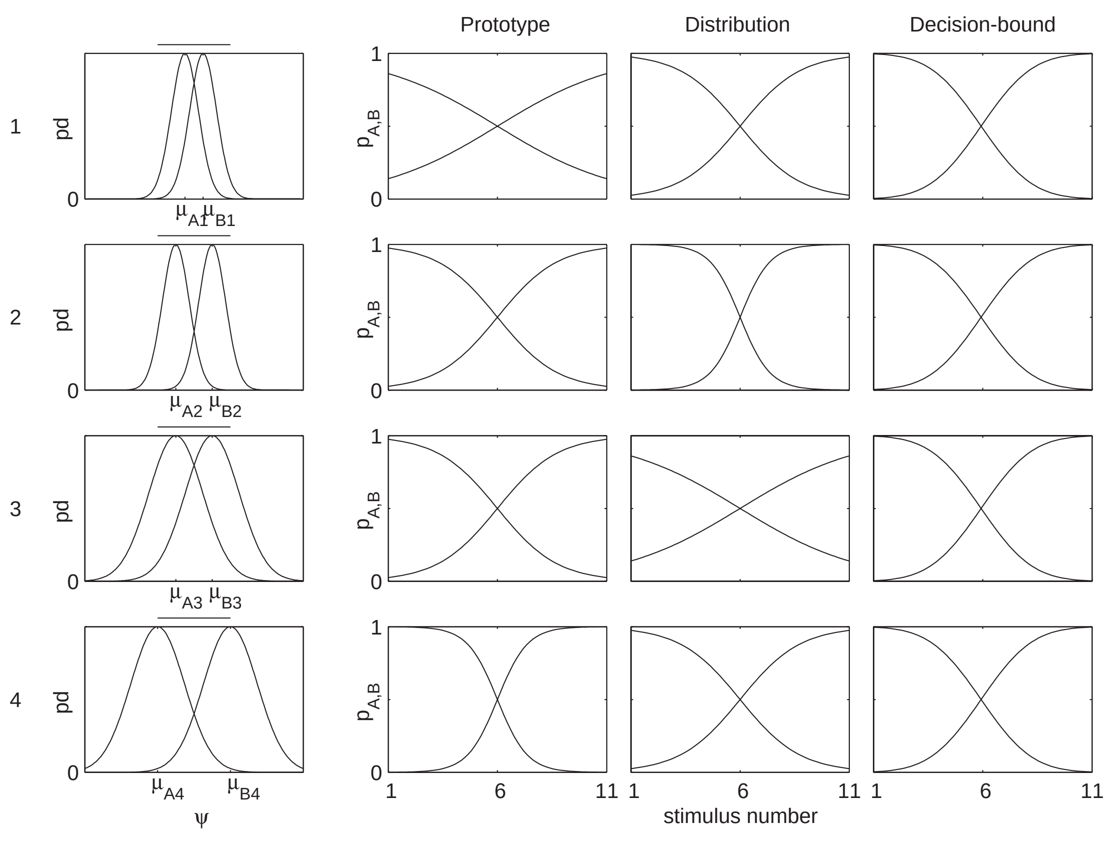

Figure 1. Training probability density functions (pdfs) and predicted categorization functions for the four experimental conditions of Experiments 1 and 3. The four panels in the leftmost column represent probability density (pd) on psychological dimension ψ in Conditions 1 through 4. μ_{Ci} Ci indicates the mean of the pdf for Category *C* in Condition *i* (e.g., μ_{B2} indicates the mean of the pdf for Category *B* in Condition 2). The small horizontal lines above the panels represent the width of the test continua. The panels in Columns 2, 3, and 4 indicate categorization functions on testing, as predicted by the prototype, distribution, and decision-bound theories of categorization. $p_{A,B}$ is short for $p(A|S_i)$ and $P(B|S_i)$.

Table 1
Training Distributions of Experiment 1

Condition	$\Delta\mu$ (jnds)	σ (jnds)	Max rate (%)	Slope ratio		
				Prototype	Distribution	Decision-bound
1	5	3.704	75.45	1	2	1
2	10	3.704	91.82	2	4	1
3	10	7.407	75.45	2	1	1
4	20	7.407	91.82	4	2	1

Note. Columns 2 and 3 give the distances $\Delta\mu$ between the means of the two training probability density functions and their standard deviations σ expressed in the number of just-noticeable differences (jnds). Column 4 gives the theoretically maximum classification rates (max rate) of an optimal, noise-free classifier for Training Conditions 1 to 4. Columns 5, 6, and 7 give the predicted ratios of the categorization function slopes in the four conditions for the prototype-, distribution-, and decision-bound–based categorization theories.

The second feature of our method was that in the test phase we used a stimulus continuum to scan participants' categorization across a relevant section of the psychological dimension ψ under study. The test phase was therefore similar to the common phonetic categorization experiment using a phonetic continuum. In the test phase, we used the same stimulus continuum across all four training conditions. Thus, any differences in the resulting categorization functions in the four conditions would be due to differences in training only. The test continua consisted of 11 stimuli with equidistant parameter values, whose lowest and highest values coincided with means μ_A and μ_B in Condition 4. The combination of the four distribution-based training conditions and the subsequent fixed test continuum allowed us to experimentally distinguish among the categorization theories. As we show, the theories predict different patterns of categorization-function slopes across the four conditions.

Theoretical Predictions

Prototype Theory

According to the prototype theory, the only information about the categories that is stored is the location of the category prototypes. If we assume Gaussian similarity functions (e.g., Nosofsky, 1986), the probability $p(A|S_i)$ of assigning stimulus S_i, defined by parameter value ψ_i, to category A is a logistic function (for mathematical derivations, see the Appendix). The slope s of this logistic function is proportional to the distance between the means of the pdfs used in the training phase. Therefore, the prototype theory predicts that the slopes of the categorization functions in Conditions 2 and 3 will be twice the slope in Condition 1, whereas the slope in Condition 4 will be four times bigger than the slope in Condition 1—that is, $s_4 = 2s_3 = 2s_2 = 4s_1$. This is graphically represented in the second column of Figure 1. The top panel, associated with Condition 1, has the shallowest categorization curves, and the bottom panel (Condition 4) has the steepest. Conditions 2 and 3 have equal slopes of intermediate values.

Distribution Theory

The distribution theory assumes that participants' category representations include not only category means but also measures of spread. When the category distributions are approximately normal, the theory assumes that participants model the categories by normal distributions, estimating for each category a mean and a standard deviation (for the unidimensional case). As is the case for the prototype theory, $p(A|S_i)$ is a logistic function of ψ_i (see the Appendix). The categorization function's slope s is proportional to the distance between the means of the training pdfs divided by their variance. As a result, Condition 3 is predicted to have the shallowest categorization functions, whereas Condition 2 will have the steepest, with a slope that is four times that for Condition 3. The categorization functions for Conditions 1 and 4 are predicted to be identical, with slopes that are two times bigger than the slope in Condition 3. In short, $s_2 = 2s_1 = 2s_4 = 4s_3$.

Exemplar Theory

The exemplar theory claims that categories are represented by the complete set of training items, whereas in the distribution theory the categories are parametric abstractions of the training items. Both theories assume a response selection mechanism on the basis of a relative goodness rule, although some recent versions of exemplar models used a deterministic response rule (Nosofsky & Zaki, 2002). Irrespective of the values of sensitivity parameter k and the Minkowski metric (the power used in the distance function; see, e.g., Ashby & Maddox, 1993) in the exemplar theory, the predicted qualitative pattern of slopes across the four conditions for the exemplar theory is expected to be identical to the pattern predicted by the distribution theory: smallest slope in Condition 3, largest in Condition 2, and intermediate in Conditions 1 and 4. Consequently, we expect the predicted response patterns for the distribution and exemplar theories to be so similar that we cannot distinguish them experimentally in the present set of experiments. Henceforth, we therefore pool the distribution and exemplar theories under the term *distribution theory*. We note, however, that, over the years, the exemplar model has been implemented in a variety of ways. Some of these implementations might lead to behavior that is different from that of the distribution models used in this article.

Decision-Bound Theory

Finally, the predictions for the decision-bound theory are straightforward. The theory assumes that, during training, participants learn the position of the optimal boundary between Categories A and B. They subsequently use this boundary in the categorization of the test stimuli. If the psychological effect of a stimulus falls to the left of the boundary, the stimulus is labeled *A*; otherwise it is labeled *B*.

Under the decision-bound theory, the slopes of the categorization functions are determined by perceptual noise only. As the test phase is identical for the four conditions, the perceptual noise is also identical, which leads to the prediction that the slopes of the categorization functions are equal across the four conditions. Note that the decision-bound theory predicts cumulative normal (probit) categorization functions rather than the logistic ones predicted by prototype and distribution theories. The two are, however, very similar and difficult to distinguish experimentally. The rightmost column of Figure 1 gives the predicted categorization functions for the four conditions, with the assumption that the standard deviation of the pdf associated with the perceptual noise equals 3.704 jnds in all four conditions.

Stimulus Considerations

We applied the experimental paradigm we have defined to two auditory dimensions that are known to be of major importance in the categorization of speech sounds: frequency of a spectral prominence, or *formant*, and duration. For example, in the categorization of the English vowels /ɛ/ and /ae/, as in the words *bed* and *bad*, respectively, both vowel duration and frequency of the first formant $F1$ are known to play a role, with /ɛ/ having shorter duration and lower $F1$ than /ae/ (e.g., Mermelstein, 1978; Whalen, 1989).

Although we expressly used speechlike dimensions in our stimuli, we endeavored to prevent the participants from explicitly using speech sounds as reference categories. The reason for this is that the experimental paradigm for distinguishing among the various theories was based on the systematic variation of the training distributions. If participants nevertheless adopted categorization strategies involving speech categories ("Respond *A* if it sounds like /ɛ/ and *B* if it sounds like /ae/"), our experiments would not measure what they were intended to measure.

We solved this problem by using a synthetic inharmonic tone complex as the base signal from which we derived the experimental stimuli. The inharmonic base signal sounded very different from speech, whose source is a mixture of a harmonic signal and noise. After taking the experiment, participants typically described the sounds as computer sounds, organs, or horns.

We created the experimental stimuli by filtering the base signal, thus creating the spectral prominence, and truncating the filtered signal to a desired duration. In Experiments 1 and 2 we varied the frequency of the spectral prominence, with all stimuli having the same duration. In Experiments 3 and 4 we varied the duration of the stimuli, keeping the formant frequency constant.

Experiment 1

Method

Participants. We recruited 67 students at Nijmegen University, Nijmegen, the Netherlands, as participants for Experiment 1. All reported normal hearing and had Dutch as their native language.

Stimuli. As we have mentioned, we derived all stimuli from a single "base signal." We constructed this base signal by adding sinusoids with exponentially spaced frequencies. The base signal $B(t)$ is defined by

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

where A is a constant amplitude factor, $f_0 = 500$ Hz is the frequency of the lowest partial, $F = 1.2$ is the frequency ratio of two successive partials, t represents time, and $N = 17$ is the number of partials. The 17 partials constituting the base signal spanned a frequency range from 500 Hz to 4679 Hz.

Next, we filtered the base signal by a single resonance or formant, implemented as a second-order infinite impulse response filter. The bandwidth of the filter was .2 times the filter's resonance frequency. Finally, we truncated the stimulus to the desired duration, applying linear 5-ms ramps at onset and offset to avoid clicks.

In Experiment 1, we varied the frequency of the formant and kept stimulus duration constant at 150 ms. Perceptual representation of frequencies, be they pure tone frequencies or formant frequencies, is often modeled by the Equivalent Rectangular Bandwidth (ERB) scale (Glasberg & Moore, 1990). The ERB scale, obtained through detailed psychoacoustic experiments, is designed such that pure tones differing by a fixed number of ERBs produce excitation patterns whose maxima have a fixed distance along the basilar membrane. We accordingly applied the earlier defined training–testing scheme (see Figure 1) to the formant frequency expressed in ERBs. We chose to vary the formant frequency roughly within the natural region of the second formant in speech.

On the basis of formant frequency discrimination data for isolated stationary vowels, Kewley-Port and Watson (1994) estimated the Weber fraction for discrimination of formant frequencies at .015 in the frequency region of the second formant. From this, it follows that at 1500 Hz, 1 jnd corresponds to 23 Hz, or .12 ERB. Using the jnd of .12 ERB for formant frequency and the earlier defined pdf means and standard deviations expressed in jnds (see Table 1), we defined pdf_A and pdf_B for the four training conditions along the ERB axis, with midpoint $\frac{1}{2}(\mu_A + \mu_B)$ at 18.7 ERB, which corresponds to 1500 Hz. Table 2 lists the resulting means and standard deviations, expressed in ERB and Hz, for pdf_A and pdf_B in Conditions 1 through 4.

The stimulus continuum for the test phase contained 11 stimuli. We obtained the formant frequencies of these stimuli by equidistant sampling of the ERB scale across the interval $[\mu_{A4}, \mu_{B4}]$ (means of pdf_A and pdf_B in Training Condition 4), which resulted in the following formant frequencies: 1288, 1329, 1370, 1412, 1455, 1500, 1546, 1593, 1641, 1690, and 1741 Hz. We used the same test continuum in all four experimental conditions.

To estimate the discriminability of the stimuli in the test continuum, we carried out an AX (same–different) discrimination experiment. This experiment is described in the Appendix. The results showed that the average discriminability of two consecutive stimuli corresponded to a d' of 1.0. This value was constant across the stimulus continuum, except for the pair 7–9, which had a higher d' than the other pairs.

Procedure. All participants first completed a training phase, after which they entered the test phase. In the training phase, two blocks of stimuli were presented, each containing all 220 training stimuli in a different randomized order. Different randomizations were used for different participants. Participants were seated in a soundproof booth in front of a computer screen. They were asked to assign sounds to either of two categories. On a given trial, a stimulus was presented binaurally through Sennheiser headphones, after which the participant categorized the stimulus by pressing either of two response buttons, labeled *A* and *B*. After the button press, the correct response was shown on the screen for 800 ms. The next stimulus was presented 700 ms after offset of the visual feedback.

Table 2
Means and Standard Deviations of pdf_A and pdf_B in the Four Training Conditions of Experiment 1 (Formant Frequency Categorization)

Condition	μ_A(ERB)	μ_B(ERB)	σ(ERB)	μ_A(Hz)	μ_B(Hz)	σ_A(Hz)	σ_B(Hz)
1	18.49	19.10	.44	1446	1559	78	84
2	18.19	19.40	.44	1393	1619	76	87
3	18.19	19.40	.87	1398	1625	153	174
4	17.58	20.01	.87	1295	1750	143	186

Note. Standard deviations in σ(ERB) column hold for both probability density functions (pdfs). ERB = equivalent rectangular bandwidth.

Before the start of training, participants were told that it would be impossible to score 100% correct, even toward the end of the training phase. Training was preceded by five familiarization trials involving stimuli drawn randomly from the 220 training stimuli. The task for the participant was the same.

The training phase was followed by a short break, after which participants entered the test phase. In this phase, participants were presented with five blocks of stimuli, each containing a different randomized ordering of four repetitions of each of the 11 test stimuli. Participants were asked to respond as quickly as possible without sacrificing accuracy. The next stimulus was played 1.5 s after a button press. No feedback was given on the correct response. After completing the experiment, participants filled out a short questionnaire asking them (a) to describe their categorization strategy, (b) whether the sounds were similar to any sound they knew, and (c) whether they thought the stimuli sounded at all like speech sounds. The entire experiment (training and test phase) took approximately 40 min.

Results

Training. Pilot experiments indicated that a fraction of the listeners had not grasped the task after completion of the training phase. We defined the following objective criterion for eliminating the data of such participants from the data set: We used a participant's data for further analysis only if he or she got at least 34 out of the last 55 training stimuli correct (corresponding to performing above chance at the $p = .05$ level). Nineteen out of 67 participants did not pass the training criterion, and we discarded their results. This left 12 participants per condition who did pass the criterion. For each participant, we divided the training data into eight consecutive blocks of 55 trials and calculated performance (percentage correct) for each block. Average performance for Blocks 1 through 8 in each of the four training conditions is plotted in Figure 2.

Figure 2 shows evidence of learning over the course of training. The average improvement from the first to the last training blocks was 15.3 percentage points. An analysis of variance (ANOVA; $MSE = 178.00$) on the difference in performance for Blocks 8 and 1, with condition as the independent variable, showed that this improvement was significant, $F(1, 44) = 63.6$, $p < .0005$, $\eta^2 = .59$. There was no significant effect of condition, $F(3, 44) = 1.7$, so the improvement was equal across the four conditions.

Participants picked up on the task reasonably quickly. Average performance during the first training block was already 10.6 percentage points above chance level (50%). An ANOVA ($MSE = 193.00$) on the difference between performance in Block 1 and chance level, with condition as the independent variable, showed this difference to be significant, $F(1, 44) = 28.2$, $p < .0005$, $\eta^2 = .39$. Condition did not have an effect on this measure, $F(3, 44) = 2.0$.

Finally, participants performed significantly below theoretically optimal performance (TOP; 75.5% in Conditions 1 and 3, 91.8% in Conditions 2 and 4) during the final training block. This was shown by an ANOVA ($MSE = 28.60$) on the difference between TOP and Block 8 performance, $F(1, 44) = 97.3$, $p < .0005$, $\eta^2 = .69$; condition had a significant effect, $F(3, 44) = 4.2$, $p = .01$, $\eta^2 = .22$. A post hoc Student–Newman–Keuls test showed that performance was further removed from TOP in Condition 4 compared with Conditions 1 and 3. However, the average deviation from optimal performance was only 7.6% and was probably mainly due to noise in the categorization process rather than premature termination of training.

Testing. The four panels in Figure 3 present categorization functions of all 12 participants for the test continuum in each of the four experimental conditions of Experiment 1. Figure 3 leaves no doubt that all participants had learned how to do the task. Stimuli with low formant frequencies (low stimulus numbers) were given predominantly *A* responses, whereas *B* responses were preferred for stimuli with high formant frequencies (high stimulus numbers). Figure 3 also suggests that all participants behaved very similarly, both within and across conditions.

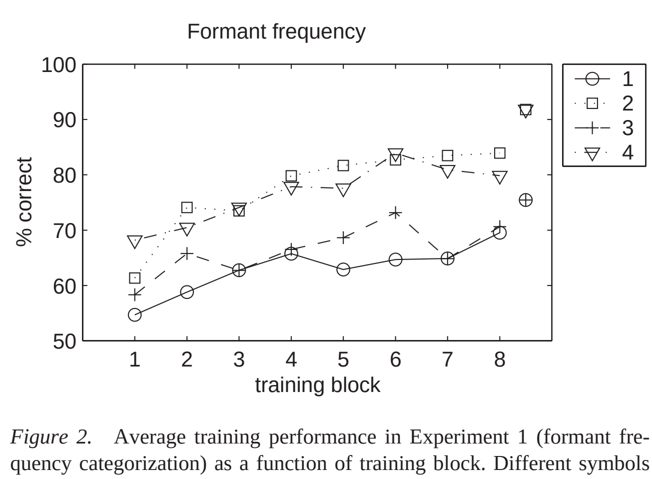

Figure 2. Average training performance in Experiment 1 (formant frequency categorization) as a function of training block. Different symbols and line types refer to different experimental conditions (Conditions 1–4). The isolated symbols to the right of Block 8 indicate theoretically optimal performance.

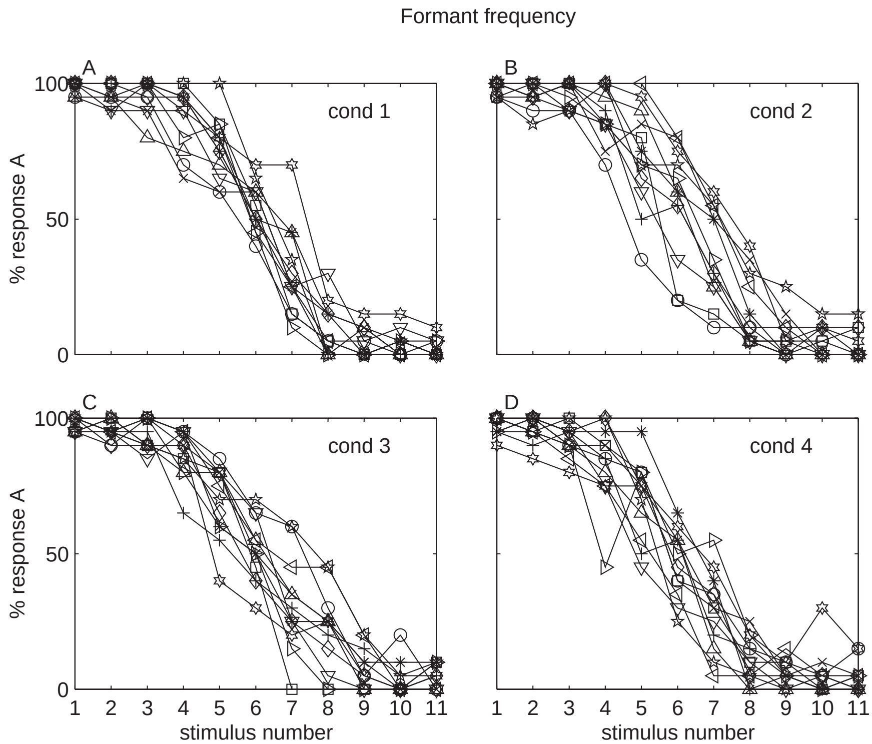

Figure 3. Categorization functions of individual participants in the four experimental conditions of Experiment 1 (formant frequency categorization). cond = condition.

To test for differences in categorization-function slopes between conditions, we carried out the following analyses. First, we performed logistic regression (e.g., Agresti, 1990) analyses on the data of each participant. In these analyses, we fitted the following model to the data of each participant.

$$\ln \frac{p(A|\psi_i)}{p(B|\psi_i)} = s(\psi_i - M), \tag{2}$$

where $p(A|\psi_i)$ is the probability of responding A to stimulus S_i characterized by value ψ_i of the perceptual representation (in this case, ERB rate) of the relevant stimulus parameter (formant frequency). s and M represent the slope and midpoint of the categorization function, respectively. We fitted Model 2 to the data of each participant, minimizing the deviance G^2 (Agresti, 1990).

The logistic regression analyses produced a slope s and a midpoint M for each participant. To test for differences in mean categorization function slopes in the four conditions, we carried out a one-way ANOVA on the slope dependent variable with condition as the independent variable. The ANOVA ($MSE = 0.10$) showed that the mean categorization slopes in Conditions 1 through 4 were not significantly different, $F(3, 44) = 1.7$, $\eta^2 = .11$. Because the prototype theory predicted a slope ratio of 4.00 between Conditions 4 and 1, we carried out an ANOVA directly comparing the slopes of these two conditions. We found no significant difference, $F(1, 22) = 1.3$, $MSE = 0.14$, $\eta^2 = .054$. The ratio of the mean slopes of Conditions 4 and 1 was 0.85, deviating strongly from the ratio of 4.00 predicted by the prototype theory. Distribution theory predicted a ratio of 4.00 between the slopes in Conditions 2 and 3. An ANOVA ($MSE = 0.05$) comparing the slopes of Conditions 2 and 3 yielded a marginally significant result, $F(1, 22) = 3.8$, $p = .065$, $\eta^2 = .15$. The ratio of the mean slopes for Conditions 2 and 3 was 1.20.

These results do not provide conclusive support for any theory. The nonsignificance of the differences among the mean slopes of all four conditions is in agreement with decision-bound theory, but this is a null result. The marginal significance of the difference between the mean slopes of Condition 2 and 3 gives partial support for the distribution theory, but the experimental slope ratio of 1.20 strongly deviates from the expected ratio of 4.00. The only firm conclusion we can make on the basis of these results is that they are in disagreement with prototype theory.

Questionnaire. Of the 48 participants who passed the training criterion, 46 described their categorization strategy essentially as “Choose *A* if the sound is low/dull; choose *B* if it is high/sharp.” The remaining 2 participants both described their strategy as “Choose *A* if it sounds like ‘*oh*’; choose *B* if it sounds like ‘*eh*.’” We compared the results of these participants during the test phase with those of other participants in the same condition, and they were very similar.

In response to Question 2, more than half of the participants said that the sounds did not remind them of any sound they knew. Typical answers of the other participants were “computer sounds,” “organ,” and “horn.” Apart from the 2 participants we have discussed, nobody mentioned speech sounds, phonemes, vowels, or anything similar in their answers to Questions 1 or 2.

When, finally, participants were explicitly asked whether they thought the sounds were speechlike (Question 3), 34 out of 48 responded *no*. One of the remaining 14 said all sounds were like "*aa*," 1 said all were like "*ee*," and the other 12 mentioned that Category A sounded like "*oh*" or a similar vowel and Category B sounded like "*ih*" or a similar vowel but added that this similarity had not occurred to them until they were explicitly asked in Question 3 (except for the 2 participants who had already reported the similarity to speech in Question 1). On the basis of these results, we concluded that we had generally been successful in preventing participants from using speech sounds as reference categories in Experiment 1.

Discussion

The results of Experiment 1 fully contradict prototype theory and give partial support for decision-bound theory and distribution theory. Although the support for decision-bound theory seems strongest, it is based essentially on a null result. In pursuit of positive effects in support of the decision-bound theory, we ran a second experiment in which the training was identical to that of Experiment 1 but in which the test continua were changed.

Experiment 2

The rationale of Experiment 2 is based on Durlach and Braida's (1969) theory of perceptual noise in decision-bound theory. Durlach and Braida hypothesized that the total variance of perceptual noise has three components: sensory variance associated with irreducible sensory (neural) noise, trace variance associated with comparisons of two consecutive sounds (as in discrimination tasks), and context variance associated with the noisy comparison of a stimulus with "perceptual anchors"—for example, the edges of the continuum used in a categorization experiment. In identification and categorization tasks, trace variance is assumed to be zero, in which case the total perceptual variance (σ^2) is the sum of sensory and context variance:

$$\sigma^2 = \beta^2 + H^2W^2, \tag{3}$$

where β^2 is the sensory variance, H is a constant, and W is the width of the test continuum expressed in psychophysical units. Equation 3 predicts that if W is small—that is, in the order of magnitude of a few jnds—context noise is small, and therefore perceptual noise is dominated by sensory noise. If, conversely, W is large—that is, the continuum spans many jnds—context noise dominates.

In our experiments, we aimed to "sample" a one-dimensional psychophysical space using a test continuum. As long as context noise is negligible, the variance of the perceptual noise is not influenced by the width of the test continuum. Consequently, if we made the width of the continuum progressively smaller, the resulting categorization function would become increasingly shallow. For example, if a given width W_1 produced a categorization function that ran from 25% to 75%, a test continuum width $W_2 = 0.5\ W_1$ would produce a shallower function running from, roughly, 37% to 63%. In conclusion, as long as W is small, the categorization function $P(A|S_i)$—the probability of choosing A as a function of stimulus number i—depends on the value of W, with a smaller W leading to a shallower function.

In the other extreme case, when W is large, perceptual noise is dominated by context noise, and sensory noise is negligible. In this case, halving the width of the test continuum will also halve the standard deviation of the perceptual noise. Paradoxically, the halved test continuum will sample an area with "halved" noise, and the categorization function $P(A|S_i)$ will be unaltered. Thus, as long as W is large, $P(A|S_i)$ does not depend on the value of W.

Imagine that, over the course of many experimental sessions, we sampled the same one-dimensional perceptual space using a set of test continua with widths ranging from close to zero to many jnds. For the very small width, the resulting categorization function would be basically flat, and with increasing W, the categorization function would become steeper. However, rather than becoming "infinitely" steep for a very large W, the slope would approach a certain asymptotic value, which we could not transgress by further increasing W. In the context of the present experiments, it is unclear where we are, exactly, on the scale of dominant sensory noise to dominant context noise. However, as long as sensory noise plays a significant role, we can use a manipulation of the width of the test continuum to produce a positive effect of experimental condition on the slope of $P(A|S_i)$.

Figure 4 presents theoretical categorization functions $P(A|S_i)$ for Experiment 2, as predicted by the three categorization theories. The two conditions of Experiment 2 are new versions of Conditions 2 and 3 in the previous experiment and are indicated as Conditions 5 and 6. The test continuum in Condition 5 was half as wide as it was in Condition 2, whereas in Condition 6 it was twice as wide as it was in Condition 3, as indicated by the horizontal bars in the leftmost panels of Figure 4.

Given our choice of test continuum widths in the new experiment, prototype theory (second column in Figure 4) predicts categorization function slopes in Conditions 5 and 6 to be identical to those in the old Conditions 1 and 4, respectively: $s_4 = s_6 = 4s_5 = 4s_1$. As we have mentioned, distribution theory (third column) predicts equal categorization slopes in Conditions 1, 4, 5, and 6. For decision-bound theory, we can make only a qualitative prediction at this point, because we do not know the value of constant H. The prediction is $s_5 < s_1 = s_4 < s_6$. For the purpose of Figure 4 (fourth column), we arbitrarily assume that sensory noise has a variance equal to variance σ^2 of the stimulus distributions in Condition 1, whereas we assumed context variance to equal σ^2 in Conditions 1 and 4, $\frac{1}{4}\sigma^2$ in Condition 5, and $4\sigma^2$ in Condition 6.

Method

Experiment 2 consists of two conditions, indicated as Conditions 5 and 6. Conditions 5 and 6 used the same training as that of Conditions 2 and 3 of Experiment 1, respectively. The width of the test continuum in Condition 5 was half the original width, covering the interval $[\mu_{A2}, \mu_{B2}]$. In Condition 6, the test continuum was twice as wide as the original one, covering the interval $[1\frac{1}{2}\mu_{A4} - \frac{1}{2}\mu_{B4}, 1\frac{1}{2}\mu_{B4} - \frac{1}{2}\mu_{A4}]$. In all cases, the number of stimuli in the test continuum was 11, as before.

Participants. We recruited 32 students at Nijmegen University as participants for Experiment 2. All reported normal hearing and had Dutch as their native language.

Stimuli. The training stimuli of Conditions 5 and 6 were identical to those of Conditions 2 and 3 of Experiment 1, respectively. The test

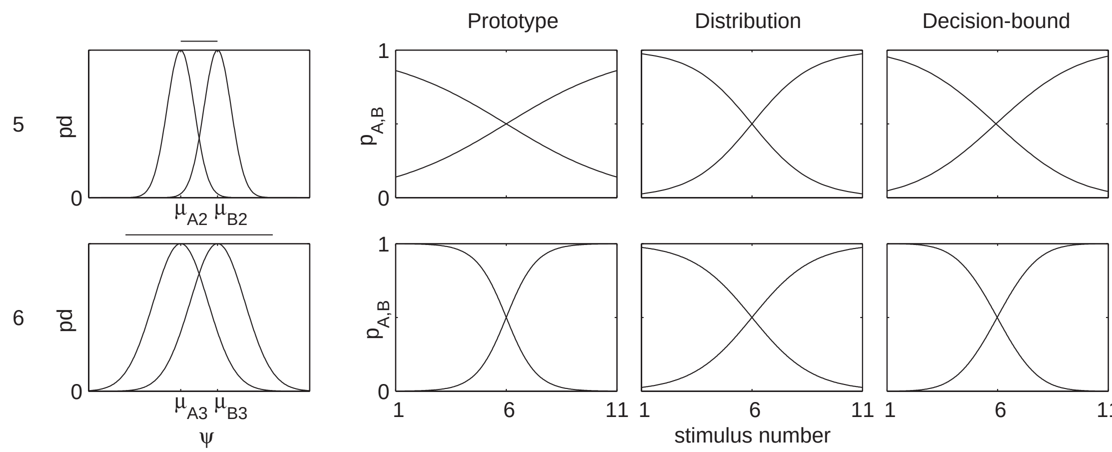

Figure 4. Training probability-density (pd) functions and predicted categorization functions for Conditions 5 and 6 of Experiments 2 and 4. The two panels in the leftmost column represent pd on psychological dimension ψ in Conditions 5 and 6. μ_{Ci} Ci indicates the mean of the pdf for Category C in Condition i (e.g., μ_{B2} indicates the mean of the pdf for Category B in Condition 2). The small horizontal lines above the panels represent the width of the test continua. The panels in Columns 2, 3, and 4 indicate categorization functions on testing, as predicted by the prototype, distribution, and decision-bound theories of categorization. $p_{A,B}$ is short for $p(A|S_i)$ and $P(B|S_i)$.

continua of Conditions 5 and 6 both contained 11 stimuli. We obtained the formant frequencies of the stimuli for Condition 5 by equidistant sampling of the ERB scale across the interval [μ_{A2}, μ_{B2}], which resulted in the following formant frequencies: 1391, 1412, 1434, 1455, 1478, 1500, 1523, 1546, 1569, 1593, and 1617 Hz. For Condition 6, we did the equidistant sampling on the interval [$1½\mu_{A4} - ½\mu_{B4}$, $1½\mu_{B4} - ½\mu_{A4}$], which resulted in the following formant frequencies: 1103, 1174, 1249, 1329, 1412, 1500, 1593, 1690, 1793, 1902, and 2016 Hz.

Procedure. The procedure of Experiment 2 was identical to that of Experiment 1 except that we did not use questionnaires, because the questionnaires of Experiment 1 had convinced us that speech-based strategies were extremely rare.

Results

Training. Eight participants did not pass the training criterion. Their results were discarded. This left 12 participants in each of the two conditions of Experiment 2. Figure 5 presents average performance for Blocks 1 through 8 in Conditions 5 and 6. It is not surprising that the learning curves of Figure 5 are similar to those of Figure 2. The average improvement from the first to the last training block in Conditions 5 and 6 was 12.1%. An ANOVA ($MSE = 47.30$) on the difference in performance for Blocks 8 and 1, with condition as the independent variable, showed that this improvement was significant, $F(1, 22) = 74.5$, $p < .0005$, $\eta^2 = .77$. There was no significant effect of condition, $F(1, 22) = 0.01$, so the improvement was equal for the two conditions.

Again, learning started quickly. Average performance during the first training block was 13.5 percentage points above chance level. An ANOVA ($MSE = 49.90$) on the difference between performance in Block 1 and chance level, with condition as the independent variable, showed this difference to be significant, $F(1, 22) = 87.4$, $p < .0005$, $\eta^2 = .80$. Condition had a significant effect on this measure, $F(1, 22) = 13.6$, $p = .001$, $\eta^2 = .38$. It is not surprising that Block 1 performance was better for Condition 5 than for Condition 6.

Finally, participants again performed significantly below TOP (91.8% in Condition 5, and 75.5% in Condition 6) during the final training block, as shown by an ANOVA ($MSE = 44.90$) on the difference between TOP and Block 8 performance, $F(1, 22) = 34.2$, $p < .0005$, $\eta^2 = .61$; condition had a significant effect, $F(1, 22) = 5.0$, $p = .04$, $\eta^2 = .18$. Performance in the final block was closer to TOP in Condition 6 than in Condition 5. However, average performance in the final block was only 8.0%, below TOP.

Testing. The two panels in Figure 6 present categorization functions $P(A|S_i)$ of all 12 individual participants for the test

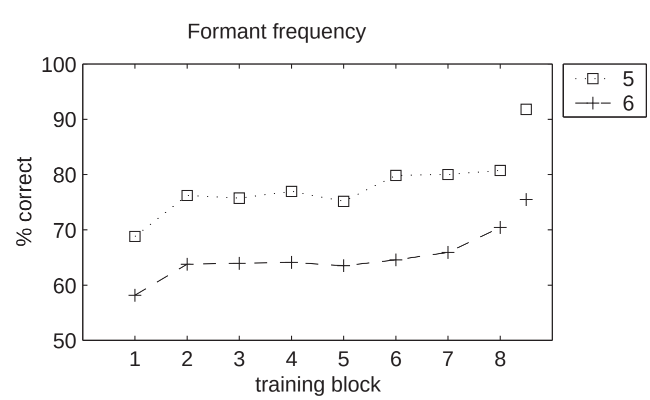

Figure 5. Average training performance in Conditions 5 and 6 of Experiment 2 (formant frequency) as a function of training block. Different symbols and line types refer to different experimental conditions (see legend). The isolated symbols to the right of Block 8 indicate theoretically optimal performance.

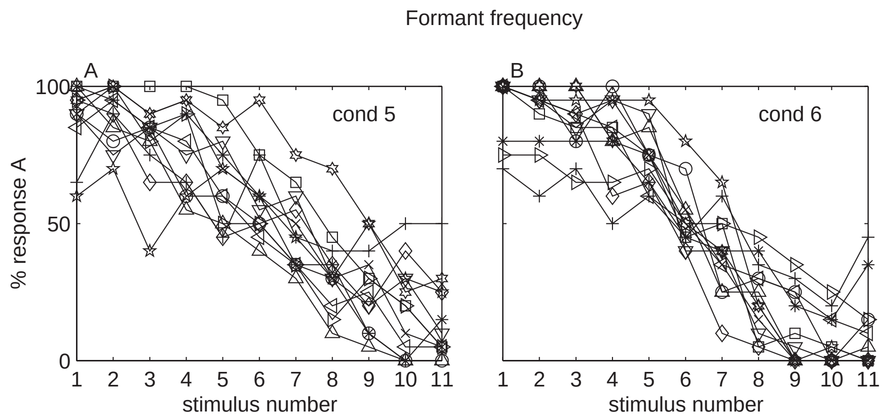

Figure 6. Categorization functions of individual participants in Conditions 5 and 6 of Experiment 2 (formant frequency). cond = condition.

continuum in Conditions 5 and 6 of Experiment 2. Comparison of Figure 6 with Figure 4 reveals that listeners' performance was more variable in Conditions 5 and 6 than in Conditions 1 through 4. The heightened variability in Condition 5 is not surprising, given the smaller test continuum width. For the variability in Condition 6, however, we have no explanation. We do note, however, that most of the variability was caused by 2 participants (the plus sign and the right-pointing triangle in Figure 6, Panel B) who, for unknown reasons, deviated somewhat from the rest. We did not have any objective criterion to remove these participants.

A one-way ANOVA ($MSE = 0.08$) showed that, on average, the categorization functions were significantly steeper in Condition 6 than in Condition 5, $F(1, 22) = 4.3$, $p = .050$, $\eta^2 = .16$. This pattern of categorization-function slopes is compatible with the decision-bound theory. An ANOVA ($MSE = 0.09$) including all six conditions (Conditions 1 through 4 for Experiment 1 and Conditions 5 and 6 for Experiment 2) showed a significant effect of condition on categorization-function slope, $F(5, 66) = 6.4$, $p < .0005$, $\eta^2 = .33$. A Student–Newman–Keuls post hoc test revealed that categorization functions in Condition 5 were, on average, significantly shallower than those of Conditions 1 through 4 and that Condition 6 was shallower than Condition 1. This general pattern is in reasonable, though not perfect, agreement with the decision-bound theory.

Discussion

In Experiments 1 and 2, we ran six conditions investigating the categorization of sounds varying in formant frequency. The pattern of categorization-function slopes was in reasonable agreement with the decision-bound theory of categorization. First, we found no difference in the slopes across the four conditions of Experiment 1, as predicted by the decision-bound theory. Second, also as predicted by this theory, the slope in Condition 5 was significantly shallower than those in Conditions 1 through 4. Condition 6 proved somewhat problematic. Whereas decision-bound theory predicted the steepest slope in Condition 6, this slope was in fact not significantly different from that of any other condition except Condition 1, compared with which it was shallower.

Theoretically, the expected increase in slope for Condition 6 was smaller than the expected decrease in slope for Condition 5, because, with increasing stimulus range, the slope increased progressively less than we would expect if performance were limited by sensory noise alone. We therefore expect that a difference between Condition 5 and Conditions 1 through 4 will reach significance earlier than the difference between Condition 6 and Conditions 1 through 4. Therefore, we consider the "asymmetry" in the slope pattern not to be in disagreement with the decision-bound theory. The shallower slope of Condition 6 compared with that of Condition 1 remains in disagreement with the decision-bound theory, however. Nevertheless, the decision-bound theory explains the data better than do the rival prototype and distribution theories, although distribution theory received weak support from the marginally significant difference between the slopes of Conditions 2 and 3.

By varying the frequency of a resonance or formant, we have varied an acoustic parameter that is generally viewed as very important for speech perception. Another such parameter is duration. To be able to draw general conclusions about the categorization mechanisms underlying speech perception, we thought it necessary to test whether the decision-bound mechanism that we found for formant frequency categorization would apply to the categorization of duration. We therefore decided to run the same set of experiments again, using similar stimuli, but this time varying stimulus duration while keeping formant frequency constant.

Experiment 3

Method

The methodology of Experiment 3 was identical to that of Experiment 1, except for the stimuli.

Participants. We recruited 53 students at Nijmegen University as participants for Experiment 3. All reported normal hearing and had Dutch as their native language.

Stimuli. In Experiment 3, we varied stimulus duration and kept the frequency of the formant constant at 1500 Hz, which was the midpoint between the means μ_A and μ_B in Experiment 1. The perceptual representation of duration has received much less attention in the psychophysical literature than that of frequency. Abel (1972) investigated duration discriminability of pure tones and noise bursts as a function of duration. The study showed that, for durations from 40 ms to 640 ms, discrimination closely followed Weber's law, with a Weber fraction of approximately .10. We carried out a pilot duration categorization experiment using this Weber fraction. The results showed that the stimuli in the duration continuum were much easier to discriminate than were those in the formant frequency continuum. Further pilot experiments indicated that assuming a Weber fraction of .05 for duration discrimination resulted in comparable discriminability of the formant frequency and duration stimuli. On the basis of these results, we defined psychological duration D, expressed in unit [d], as

$$D = 10 \log T, \tag{4}$$

where T is physical duration, expressed in milliseconds. One jnd for duration corresponds to 0.5 d. Using Equation 4, we defined pdf$_A$ and pdf$_B$ for the four training conditions along the D axis, with midpoint ½(μ_A + μ_B) at 50.11 d, which corresponds to 150 ms. Table 3 lists the resulting means and standard deviations expressed in d and milliseconds for pdf$_A$ and pdf$_B$ in Conditions 1 through 4.

As in Experiment 1, the stimulus continuum for the test phase contained 11 stimuli. We obtained the durations of these stimuli by equidistant sampling of the D scale across the interval [μ_{A4}, μ_{B4}], which resulted in the following durations: 91.0, 100.5, 111.1, 122.8, 135.7, 150.0, 165.8, 183.2, 202.5, 223.8, and 247.3 ms. We used the same test continuum in all four experimental conditions.

We tested the discriminability of the duration stimuli in the same discrimination experiment as that of the formant frequency stimuli (see the Appendix). The average discriminability of two consecutive stimuli on the duration continuum corresponded to a d' of 0.8, which was not significantly different from the d' for formant frequency discrimination (1.0). Again, discriminability was constant across the test continuum.

Procedure. The procedure of Experiment 3 was identical to that of Experiment 1.

Results

Training. Four out of 53 participants did not pass the training criterion, and 1 participant responded randomly during the test phase, having misunderstood the instructions. Their results were discarded. This left 12 participants per condition. Figure 7 presents average performance across the eight training blocks.

Learning was so quick that participants were already performing close to ceiling during the first block of training. The average improvement from the first to the last training block was only 5.5 percentage points. An ANOVA (MSE = 78.40) on the performance difference for Blocks 8 and 1, with condition as the independent variable, showed that this improvement was significant,

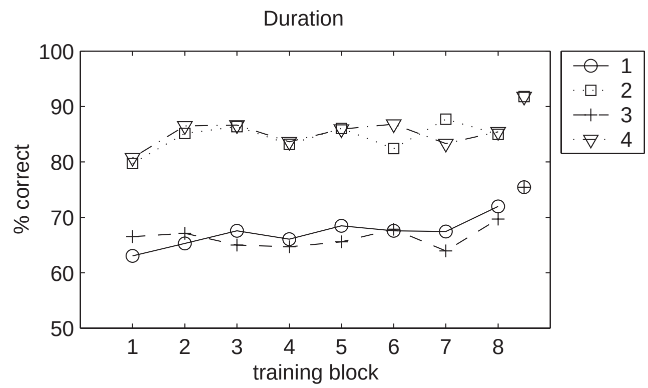

Figure 7. Average training performance in Conditions 1 through 4 of Experiment 3 (duration) as a function of training block. Different symbols and line types refer to different experimental conditions (see legend). The isolated symbols to the right of Block 8 indicate theoretically optimal performance.

$F(1, 44) = 18.7$, $p < .0005$, $\eta^2 = .30$. There was no effect of condition, $F(3, 44) = 0.9$.

Participants' learning was extremely fast. The average performance during the first training block was already 23 percentage points above chance level (50%). An ANOVA (MSE = 63.10) on the difference between performance in Block 1 and chance level, with condition as the independent variable, showed this difference to be significant, $F(1, 44) = 386.0$, $p < .0005$, $\eta^2 = .90$. Condition proved to have a significant effect on this measure, $F(3, 44) = 15.6$, $p < .0005$, $\eta^2 = .52$, and it is not surprising that a post hoc Student–Newman–Keuls test showed that Block 1 performance was higher in Conditions 2 and 4 than in the other two conditions.

Finally, participants performed significantly below TOP (75.5% in Conditions 1 and 3, 91.8% in Conditions 2 and 4) during the final training block. This was shown by an ANOVA (MSE = 26.80) on the difference between TOP and Block 8 performance, $F(1, 44) = 56.3$, $p < .0005$, $\eta^2 = .56$; condition had no significant effect, $F(3, 44) = 1.0$. The average deviation from optimal performance was only 5.6%.

Testing. The four panels in Figure 8 present categorization functions of all 12 participants for the test continuum in each of the four experimental conditions of Experiment 3. Figure 8, similar to the results for formant frequency categorization (see Figure 3), shows relatively little variability among participants within con-

Table 3
Means and Standard Deviations of pdf$_A$ and pdf$_B$ in the Four Training Conditions of Experiment 3 (Duration Categorization)

Condition	$\mu_A(d)$	$\mu_B(d)$	$\sigma(d)$	μ_A(ms)	μ_B(ms)	σ_A(ms)	σ_B(ms)
1	48.86	51.36	1.79	134.5	172.7	24.2	31.1
2	47.61	52.61	1.79	118.7	195.7	21.4	35.2
3	47.61	52.61	3.58	124.5	205.2	45.5	74.9
4	45.11	55.11	3.58	96.9	263.5	35.4	96.2

Note. Standard deviations in the $\sigma(d)$ column hold for both probability density functions (pdfs).

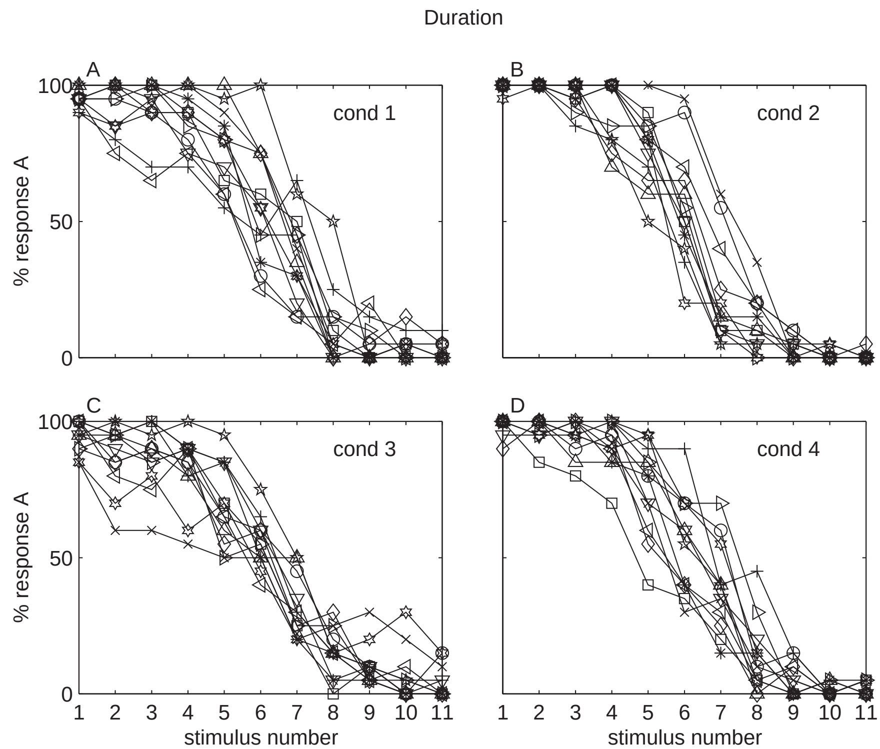

Figure 8. Categorization functions of individual participants in the four experimental conditions of Experiment 3 (duration). cond = condition.

ditions. *A* and *B* responses were preferred for short stimuli (low stimulus numbers) and long stimuli (high stimulus numbers), respectively.

As for Experiment 1, we calculated slope and midpoint for each categorization curve by means of logistic regression and subjected the slope values to a one-way ANOVA, with condition as the independent variable. The analysis ($MSE = 0.14$) showed a significant effect of condition, $F(3, 44) = 4.6$, $p = .007$, $\eta^2 = .24$. A post hoc Student–Newman–Keuls test revealed that the mean categorization function slope was significantly smaller in Condition 3 than in Condition 2, as predicted by the distribution theory. The ratio of mean slopes for Conditions 2 and 3 was 1.70, which was smaller than the predicted ratio of 4.00.

Questionnaire. All participants spontaneously described Categories A and B as short versus long, respectively. When asked what sounds the stimuli reminded them of, none of the participants mentioned speech sounds. As in Experiment 1, typical answers were computer sound, organ, and horn. When participants were explicitly asked whether the stimuli sounded at all like speech, 5 participants mentioned a single vowel, whereas 4 participants mentioned a vowel pair. From these responses, we concluded that none of the participants explicitly used a categorization strategy involving speech sounds.

Experiment 4

As with Experiment 2, we also decided to test categorization of duration by means of two continua that varied in width.

Method

The methodology of Experiment 4 was identical to that of Experiment 2, except for the stimuli.

Participants. We recruited 25 students at Nijmegen University as participants for Experiment 4. All reported normal hearing and had Dutch as their native language.

Stimuli. The training stimuli used in Conditions 5 and 6 of Experiment 4 were identical to those of Conditions 2 and 3 of Experiment 3. The 11-member test continuum of Condition 5 had a range that was half that of the test continuum of Experiment 3. The stimulus durations were 116.8, 122.8, 129.1, 135.7, 142.7, 150.0, 157.7, 165.8, 174.3, 183.2, and 192.6 ms. The 11-member test continuum of Condition 6 had twice the range of the test continuum of Experiment 3, with stimulus durations of 55.2, 67.4, 82.3, 100.5, 122.8, 150.0, 183.2, 223.8, 273.3, 333.8, and 407.7 ms.

Procedure. The procedure of Experiment 4 was identical to that of the previous experiments, except that we did not use questionnaires.

Results

Training. One of the 25 participants did not pass the training criterion. We discarded her results, which left 12 participants per condition. Figure 9 presents average performance across the eight training blocks.

As we expected, Figure 9 closely resembles Figure 7 in all respects. The average improvement from the first to the last training block was only 6.5 percentage points, which proved significant ($MSE = 118.00$), $F(1, 22) = 8.7$, $p = .008$, $\eta^2 = .28$. There was no effect of condition on this improvement, $F(1, 22) =$

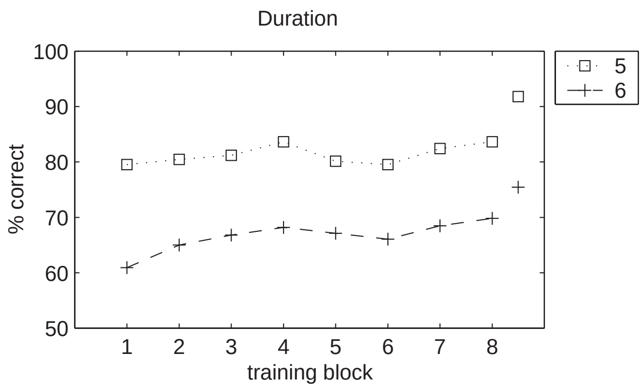

Figure 9. Average training performance in Conditions 5 and 6 of Experiment 4 (duration) as a function of training block. Different symbols and line types refer to different experimental conditions (see legend). The isolated symbols to the right of Block 8 indicate theoretically optimal performance.

1.2. Average performance during the first training block was 20 percentage points above chance level (50%). As in Experiment 3, the difference between performance in Block 1 and chance level was significant ($MSE = 85.10$), $F(1, 22) = 115.0$, $p < .0005$, $\eta^2 = .84$. This difference was significantly larger in Condition 6 than in Condition 5, $F(1, 22) = 24.4$, $p < .0005$, $\eta^2 = .53$. Also as in Experiment 2, participants performed slightly (6.9 percentage points) but significantly below TOP during the final training block ($MSE = 32.50$), $F(1, 22) = 35.1$, $p < .0005$, $\eta^2 = .62$. There was no effect of condition on this difference, however, $F(1, 22) = 1.2$.

Testing. Figure 10 presents individual categorization functions in Conditions 5 and 6 of Experiment 4. As before, we subjected the slopes of each of the individual categorization functions of Conditions 5 and 6 to a one-way ANOVA ($MSE = 0.12$) with condition as the independent variable. The analysis showed that the categorization functions in Condition 6 were significantly steeper than those in Condition 5, $F(1, 22) = 5.7$, $p = .03$, $\eta^2 = .21$. This result is in agreement with the decision-bound theory.

A combined analysis of the data of Experiments 3 and 4 (Conditions 1 through 6; $MSE = 0.13$) confirmed that mean slopes were different across conditions, $F(5, 66) = 5.5$, $p < .0005$, $\eta^2 = .29$. A post hoc Student–Newman–Keuls test revealed that the average categorization-function slope in Condition 5 was significantly shallower than the slopes in Conditions 1, 2, and 4. In addition, the slope of Condition 2 was significantly steeper than that of Condition 3. The overall pattern found for the combined analysis lies between the patterns expected for the distribution and decision-bound theories.

Discussion

Experiments 3 and 4 test the basic mechanism underlying the categorization of sounds that vary in duration. The results show evidence for two theories of categorization. In Experiment 3, we found that categorization-function slopes were larger in Condition 2 than in Condition 3, whereas the slopes for Conditions 1 and 4 were in between. This pattern of slopes is in agreement with the distribution theory. In Experiment 4, however, the slope was steeper in Condition 6 than in Condition 5, which is in agreement with the decision-bound theory. Thus, the combined results give partial support for two theories. The only theory that remains unsupported is the prototype theory.

The combined results of the four experiments raise two important questions. First, how can we interpret the partial support for two theories? A possibility is that aspects of the distribution and decision-bound theories should be combined. Second, why do we find different results for duration and formant frequency? The results for stimuli that vary in formant frequency support the decision-bound theory, whereas the results for duration are in partial agreement with the decision-bound and distribution theories. We address both questions in the next sections.

Model Analyses

The data analyses we have presented concentrate on the qualitative patterns of categorization slopes across conditions. The prototype, distribution, and decision-bound models, however, are

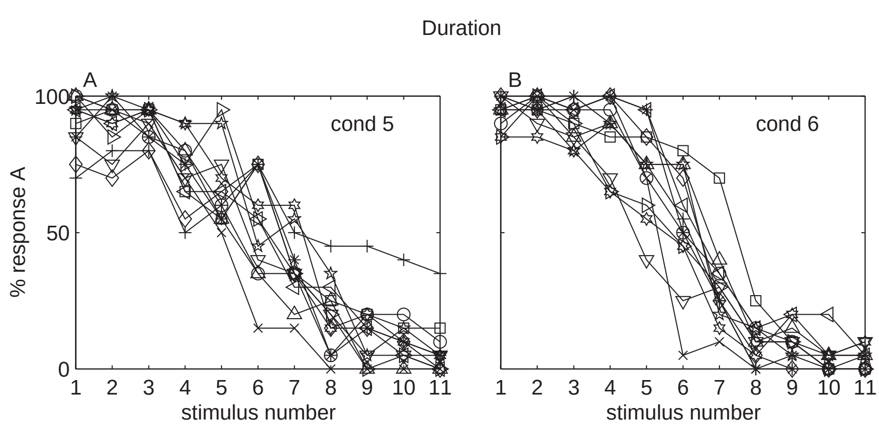

Figure 10. Categorization functions of individual participants in Conditions 5 and 6 of Experiment 4 (duration). cond = condition.

mathematically fully developed and allow for quantitative testing. Apart from a small number of free parameters, each of the models can predict quantitative data for each of the six experimental conditions used in the experiments. In particular, such quantitative analyses may shed some light on the interesting but unsatisfactory finding that the duration stimuli seem to have been categorized by a mixture of decision-bound and distribution strategies.

Method

We performed the model analyses on individual data only. The reason for this choice is that the slope of pooled categorization curves is very sensitive to variation in the curve midpoint in the individual data. That is, summing two steep categorization curves with widely separated midpoints yields a shallow categorization curve. As the categorization-function slope is the dependent parameter in our experiments, we thought we should avoid such harmful effects of data pooling. To get optimal fits, we used the individual midpoints of the categorization functions estimated by the individual logistic regressions in our model analyses. Before calculating the goodness of fit for each participant, we shifted the predicted categorization function to coincide with that participant's midpoint.

The absence of sensory noise in prototype and distribution models can be viewed as either a basic assumption or an approximation for superthreshold categorization problems. Because we designed our stimuli such that test continuum neighbors were moderately confusable, the approximation may not apply, and we have to allow for the possibility that perceptual noise played a role even if the listeners in our experiments used a prototype- or distribution-based categorization mechanism. We therefore fitted "expanded" versions of the prototype and distribution models that incorporated perceptual noise. In the case of the prototype model, we added only context noise, because the addition of sensory noise was mathematically equivalent to a decrease in the decay of similarity—that is, a lower value of parameter k.

For each class of model, we made a series of fits with increasing numbers of free parameters associated with (depending on the model class) the two kinds of perceptual noise and the decay of similarity. Using an overdispersion-based technique (e.g., McCullagh & Nelder, 1989), we determined for each model class which extra free parameters significantly improved the fit. The comparison of models of different classes was less straightforward, however. When models are not hierarchically related—that is, none of the models is a special case of any other model—formal statistical testing is impossible. Recent studies comparing nonhierarchical models have used the Akaike (1974) information criterion as a measure of goodness of fit (e.g., Ashby, Maddox, & Bohil, 2002). For repeated-measures data such as those used in our experiments, however, the Akaike information criterion is generally found to "underpunish" the addition of free parameters. In addition, any measure of goodness of fit is noisy, so if the difference in model performance is small, one should be cautious in selecting the "winning" model. We therefore based the comparisons among model classes not only on goodness of fit and number of free parameters but also on the extent to which each of the models was able to replicate global trends in the data.

We fitted all models to experimental data using a general-purpose nonlinear minimization technique. We found parameter values that minimized the deviance G^2. We made separate model fits for the formant frequency data (Experiments 1 and 2) and the duration data (Experiments 3 and 4). Given a stimulus dimension (frequency or duration), we used the simplifying assumption that a single set of parameters applied to all participants. Alternatively, one may assume that participants' parameter values were sampled from a distribution whose mean and spread are constant across conditions. Even if one knew the appropriate distribution (which we do not) this assumption would make the model-fitting procedure much more complex, and it would probably lead to the same conclusions. Therefore, we made a single fit of each model to all data for each stimulus dimension—that is, the data of all 72 participants in all six conditions.

Results and Discussion

The addition of perceptual noise to the prototype model did not significantly improve the fit for either stimulus dimension. The fit for the distribution model, conversely, improved with the addition of both sensory and context noise for formant frequency as well as duration. Removing either sensory or context noise from the decision-bound model always significantly worsened the fit.

Table 4 presents the details of the analysis results. We included the results of the distribution theory without perceptual noise in the table to allow for comparison of the three standard models. The last row, labeled *NENA*, refers to a new model, which we define in the next section. First, the results confirm that of the three models we tested, the prototype theory provided the worst account of the data. Both for the formant frequency and for the duration continua, the prototype model gave the worst fit, with G^2 values that were much larger than those for the other two models.

Compared with the standard version of the distribution model, the decision-bound model gave the best fit for both stimulus dimensions. This result is interesting because the results of Experiment 3 suggest that listeners used a distribution-based categorization method for the duration dimension. The present model analyses indicate that although the duration data show a significant qualitative pattern in accordance with the distribution theory, the

Table 4
Results of Model Analyses

	Formant frequency		Duration	
Model	Parameter values	G^2	Parameter values	G^2
Prototype	$k = 1.1$	2402	$k = .076$	2029
Distribution	$k = .91$	2124	$k = 1.1$	1577
Distribution, PN	$k = 2.0$, $\beta = .36$ ERB, $H = .10$	1884	$k = 1.9$, $\beta = 1.3\ d$, $H = .059$	1372
Decision bound	$\beta = .29$ ERB, $H = .21$	1631	$\beta = 1.1\ d$, $H = .19$	1536
NENA	$\beta = .29$ ERB, $H = .21$, $\alpha = .010$	1630	$\beta = 1.1\ d$, $H = .13$, $\alpha = .62$	1431

Note. Parameters k, β, and H model the similarity gradient, the standard deviation of the sensory noise, and the coefficient of the stimulus range in the context variance, respectively. ERB = equivalent rectangular bandwidth; NENA = noisy encoding noisy activation; PN = perceptual noise.

decision-bound theory still provided the best quantitative account of the results (although the difference was small and possibly not meaningful). However, when we augmented the distribution model to allow for perceptual noise, its fit improved significantly for both stimulus dimensions. For the formant data, G^2 was reduced by 11%. Despite this reduction, the augmented distribution model still did worse than the decision-bound model. For the duration data, conversely, G^2 was reduced by 13%, and the resulting fit was better than that of the decision-bound model.

Recall that Equation 3 expresses the decomposition of the total variance of the perceptual noise into two components: sensory variance (β^2) and context variance (H^2W^2). The fourth row of Table 4 gives the values of β in the decision-bound models for the two stimulus dimensions, expressed in psychophysical units (ERB and d, respectively). If we convert these into number of stimulus steps (on the continua of Conditions 1 through 4), we find that the βs for the formant frequency and duration dimensions are 1.2 and 1.1 stimulus steps, respectively. The similarity of the two values shows that the step sizes we used for the continua on the two stimulus dimensions were well chosen, because the discriminabilities of successive steps on the two continua are comparable.

For the manipulation of Experiment 2 to yield a positive prediction for the decision-bound models, sensory variance needed to be nonnegligible. Using the estimated parameter values, we can check whether we chose the test continuum widths appropriately. The values of the coefficient H for formant frequency and duration were .21 and .19, respectively. From these values and the values for β, we estimate the ratio of the standard deviations of context and sensory noise at 1.8 for the formant frequency dimension and 1.7 for the duration dimension. On the one hand, these values show that sensory variance and context variance were in the same order of magnitude for both stimulus dimensions and that it was reasonable to expect a measurable difference in categorization-function slopes for Conditions 5 and 6. On the other hand, because the contribution of context noise is bigger than that of sensory noise, the slope difference should not be very large.

We now turn to the parameter estimates for the augmented distribution models. For both stimulus dimensions, the power parameter k equals roughly 2. This indicates that similarity does not decay proportionally to the category likelihood but decays faster, namely proportionally to the squared likelihood. If we compare the parameters coding with the perceptual noise in the distribution models and the decision-bound models, we find that the sensory noise (parameter β) is roughly equal in the two models, whereas the context noise (H) is smaller in the distribution models than in the decision-bound models (ratio of 0.50 for formant frequency and 0.30 for duration).

We can interpret these parameter values as follows. A power parameter k of 1 is compatible with a similarity function proportional to the category likelihood, followed by a response selection process governed by the Luce (1963) choice rule. However, if k approached infinity, the simple distribution model would become a noise-free decision-bound model because the ratio of the similarities to the two categories would be either zero or infinity (e.g., Ashby & Maddox, 1993). Analogously, the augmented distribution model with infinite k would become a standard (i.e., noisy) decision-bound model and would be governed by perceptual noise and deterministic response selection. (Note, however, that a very large value of k leads to similarity values close to zero for all categories. Although such values are, of course, mathematically possible, they are conceptually incompatible with one of the core assumptions of the distribution model, which is the similarity calculation.) The parameter values show that, at least for the duration data, the truth lies somewhere in the middle. Sensory noise is roughly equal in the decision-bound models and the augmented distribution models. k in the augmented distribution models is larger than 1, so the models approach the decision-bound models somewhat. Thus, our model analyses suggest that the categorization mechanism used by our listeners has aspects of both distribution and decision-bound theories.

Figure 11 presents the mean slopes of the observed and modeled categorization functions across all six conditions of both stimulus

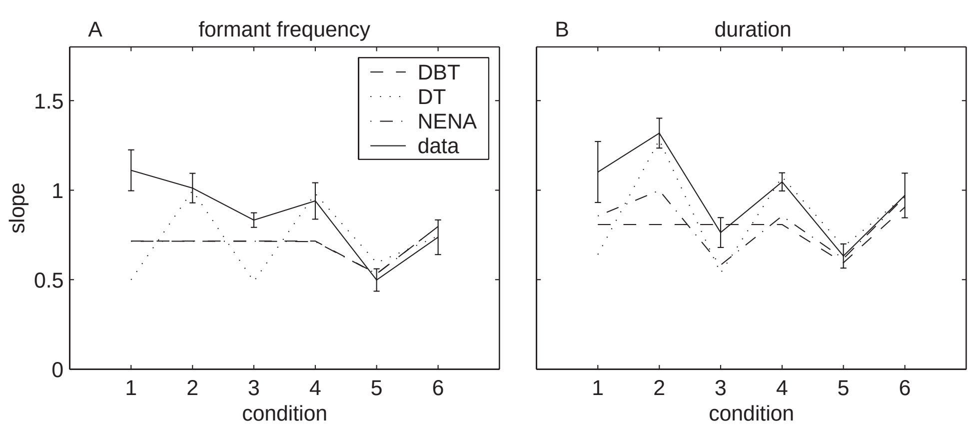

Figure 11. Predicted and observed mean categorization-function slopes in Conditions 1 to 6 for formant frequency (see Panel A) and duration (see Panel B). Solid lines give observed slopes, with vertical bars representing plus or minus one standard error. Dashed, dotted, and dashed-dotted lines give theoretical slopes, as predicted by the decision-bound theory (DBT), the distribution theory (DT), and the noisy encoding noisy activation (NENA) model, respectively.

dimensions. The solid lines give the slopes of the categorization functions derived from the experimental data. The vertical line segments indicate plus or minus one standard error. The dashed and dotted lines represent the slopes expected by the decision-bound and augmented distribution models as well as a third model that we discuss later. (Note that we optimized the models to fit the raw data, not the slope values.)

Figure 11 first reconfirms that the observed mean categorization-function slopes for the two stimulus dimensions are of the same order of magnitude. This tallies with our earlier finding that the best fitting perceptual-noise variance was of similar size for the two stimulus dimensions. Second, the figure provides us with extra means of evaluating the goodness of fit of the theories to the data, this time not the raw data but the slopes. We first focus on the formant frequency data (see Figure 11, Panel A). There are two ways in which we can judge the fit: First, we can simply examine how close the expected slopes are to the experimental ones. If we judge a model fit to be satisfactory when it falls within plus or minus two standard errors of the data, we find that neither theory on its own gives a satisfactory account of the data for either stimulus dimension. The decision-bound theory, which provided the best fit to the raw data, matched the observed slopes only in Conditions 5 and 6. The augmented distribution theory matched the observed slopes in Conditions 2, 4, 5, and 6 but deviated very strongly in Conditions 1 and 3. Second, we can judge how well each of the models replicates the overall slope patterns. For formant frequency, the overall pattern was that slopes were equal across all conditions except Condition 5, for which the slope was smaller. This pattern was replicated more closely by the decision-bound theory than by the augmented distribution theory.

For duration, the decision-bound theory matched the observed slopes in Conditions 1, 3, 5, and 6, whereas the augmented distribution theory matched the observed slopes in Conditions 2, 4, 5, and 6. Here, however, the replication of the overall slope pattern plays a decisive role in the evaluation. The decision-bound model did not and cannot replicate the overall slope pattern, whereby the slope in Condition 2 was steeper than in Condition 3. The augmented distribution theory did replicate the overall pattern, although the variation in the experimental slopes was stronger than in the actual data.

Noisy Encoding Noisy Activation (NENA): A New Model of Categorization

Figure 11 shows that neither model fits the overall slope patterns. Even the augmented distribution theory, which includes a power parameter for flexibility in the decay of similarity and incorporates both types of noise of the decision-bound theory, does not provide a satisfactory account of the categorization-function slopes across the two stimulus dimensions. In this section, we propose a new model of categorization that may explain the present results more fully. First, however, we reexamine our results in terms of the necessary components of any categorization theory—that is, stimulus encoding, category representation, and response selection.

Concerning stimulus encoding—that is, the manner in which the stimulus is mapped onto a point in perceptual space—our model analyses give strong support for an important role of perceptual noise. The decision-bound theory, in which perceptual noise plays a pivotal role, gave a superior account of the frequency data. Furthermore, the fits of the distribution model improved considerably for both stimulus dimensions when we added perceptual noise to the model. Note that, in the context of exemplar models, noisy stimulus encoding has been proposed for confusable one-dimensional stimuli (e.g., Ennis, 1988).

Concerning category representation, the results speak equally clearly. As both the distribution and the decision-bound theories support category representation in the form of distributions, our results give strong support for the distribution representation. Listeners do not just store the mean or best representative of a set of category members; they also include information on the spread of the category in their representation. We reiterate that the present research cannot decide on whether this distribution is parametric (e.g., Gaussian) or nonparametric (e.g., fully exemplar based). Concerning the decision process, our data are less clear. There is evidence of both a deterministic and a stochastic decision process. A potential approach to modeling such a mixture of processes is criterial noise. Ashby and Maddox (1993) discussed how criterial noise—that is, noise in the location of the decision bound—may be incorporated into decision-bound models. We assumed that if a decision-bound model with criterial noise applied, listeners would search during training for a boundary position that optimized their percentage correct rate. If the percentage correct rate (which is determined by the likelihood ratio of the two training distributions) changed rapidly with the boundary estimate, listeners would quickly approach an accurate boundary location, whereas if the percentage correct rate was relatively insensitive to the boundary estimate, it would take listeners a long time to find an accurate boundary location. We therefore thought it reasonable to assume that the standard deviation σ_B of the criterial noise would be inversely proportional to the change of the percentage correct rate P_c during training with changing boundary location B:

$$\sigma_B \sim \frac{dB}{dP_c}. \tag{5}$$

The proportions of $\frac{dP_c}{dB}$ for Conditions 1 through 4 were 4:2:2:1, respectively. This means that the proportions of the standard deviations of the criterial noise would be 1:2:2:4, respectively. Thus, if listeners used a decision-bound mechanism in which criterial noise played a significant role, the pattern of categorization-function slopes would be steep–intermediate–intermediate–shallow for Conditions 1 through 4, respectively—that is, the opposite of the prototype pattern. This prediction only makes the decision-bound model move away from the observed slope pattern. Therefore, the addition of criterial noise to the decision-bound model cannot explain the observed results either.

On the basis of these considerations, we constructed a new hybrid model with the following properties. Concerning category representation, we assume that categories are represented by distributions. These distributions are learned through previous exposure, as in the training phase of our experiments. The distributions are either parametric or nonparametric. For the purpose of the present study, we assume they are parametric and have the form of normal distributions characterized by a mean and variance. Concerning the processing issue, we assume that the stimulus encoding is stochastic—that is, there is perceptual noise. As in the decision-bound theory, this noise is normal, with zero mean and a variance with two components: trace variance and context variance. After stimulus encoding, the stimulus is represented as a point $\tilde{\psi}$ on a psychological axis, where the tilde ($\sim$) indicates that a noise component is present. Next, the stimulus is mapped onto category similarity (or activation) values for each of the two categories in a manner similar to that of the distribution model. In contrast to the distribution model, however, this calculation is assumed to be stochastic. First we calculate the value of the distribution at the stimulus location. We can indicate this value as $A(\tilde{\psi})$, where A represents the category distribution or *activation function* of the category. Next, we take the logarithm $\log A(\tilde{\psi})$, to which we add normal noise with mean zero, which results in a value $\widetilde{\log}\ A(\tilde{\psi})$. Next, we calculate the exponent, which results in the noisy activation value $\exp \widetilde{\log}\ A(\tilde{\psi})$ or, given more compactly, $\tilde{A}(\tilde{\psi})$. We chose to add normal noise to the logarithm of the activation instead of to the "raw" activation for two reasons. First, adding normal noise to $\log A$ leads to the Luce (1963) choice rule, whereas adding it to A does not (Albert & Chib, 1993). Second, conceptually, adding normal noise to an activation value (or similarity) is incorrect because it can lead to values below zero. As a final step, we select the category that has the largest activation value. The latter choice process is deterministic. We call the hybrid model NENA to reflect the essential components of the model.

The NENA model is a true hybrid. On the one hand, it contains perceptual noise and deterministic choice, similar to the decision-bound model. On the other hand, an essential step in the model is a category-activation calculation that compares the incoming stimulus with the distribution of training stimuli, as in the distribution model. As such, the decision-bound model does not play an explicit role in the categorization process. If the noise in the activation calculation (henceforth *activation noise*) is zero, the model is a standard decision-bound model. If the perceptual noise is zero and the shape of the activation noise is such that it leads to a response behavior that is mathematically equivalent to the Luce (1963) choice rule, the model is equivalent to the distribution model. By varying the amounts of perceptual and activation noise, we should be able to obtain satisfactory fits to the data for both dimensions using a single model.

The NENA model has three free parameters: β and H coding sensory and context noise, respectively, and α representing the standard deviation of the activation noise—that is, the Gaussian noise added to the logarithm of the category similarities. Unfortunately, analytical solutions linking stimulus values to response probabilities do not exist, so we had to resort to Monte Carlo techniques to fit the hybrid model to our data. To obtain reliable model estimations, we generated 300,000 random values for each stimulus in each of the six conditions and then used a minimization procedure to find the parameter values that produced the lowest value of G^2.

For formant frequency, the best fitting model had a G^2 equal to 1630. This value is very close to that of the decision-bound theory ($G^2 = 1631$) that we obtained earlier. Apparently, "adding" a distribution theory component to the model did not result in an improvement in goodness of fit. The similarity of the two models is further corroborated by the fact that the best fitting values of β and H of the hybrid model are identical to those for the decision-bound theory (1.2 stimulus steps and 0.21, respectively; see Table 4). The value of α is 0.010—that is, almost zero. Thus, the hybrid model does not account for the slight distribution theory–like trend in the slope data that we discussed earlier.

The best fitting NENA model for duration had a G^2 equal to 1431. The fit is therefore somewhat worse than the fit of the augmented distribution model ($G^2 = 1373$) but better than that of the decision-bound model. The best fitting values of the model parameters are $\beta = 1.1$, $H = 0.13$, and $\alpha = 0.62$. If we compare these with the values for the decision-bound model ($\beta = 1.1$, $H = 0.19$), we see that the context noise parameter has decreased. Effectively, the NENA model assigns a significant portion of the total noise in the process to the similarity calculation, leaving less for the perceptual encoding. This tallies with the finding that α is larger for duration than for formant frequency (0.62 vs. 0.01, respectively). Finally, we note that the summed G^2 for the two dimensions is smaller for the NENA model ($G^2 = 3060$) than for both the decision-bound theory ($G^2 = 3167$) and the augmented distribution theory ($G^2 = 3257$), although the advantage is small.

The dash-dotted line in Figure 11 gives the categorization function slopes of the NENA model. For formant frequency, the slope values predicted by the NENA model are not noticeably different from those predicted by the decision-bound theory. For duration, conversely, the NENA model predicts slope values that truly combine aspects of the distribution theory and the decision-bound theory. The pattern of slope values exhibits both the (weakened) distribution theory–like pattern for Conditions 1 through 4 and the context-noise effects for Conditions 5 and 6.

In sum, we have formulated a hybrid model of categorization that gives a reasonable account of our experimental results for both the formant frequency and the duration dimensions through a combination of aspects of the decision-bound and distribution theories of categorization. We do note, however, that some discrepancies between data and model predictions remain.

General Discussion

The present study addresses the question of how listeners categorize sounds. In four experiments, we trained listeners to categorize synthetic sounds from two overlapping distributions. Subsequently, listeners categorized stimuli from a test continuum without receiving feedback. The crucial manipulations in the experiments were variation of the variance and overlap of the two distributions as well as the width of the test continuum. Prototype, distribution, and decision-bound theories of categorization made different predictions about the slopes of the categorization functions in the various conditions.

When we applied the methodology to the formant frequency dimension, we found a pattern of slopes that was in reasonable agreement with the decision-bound theory. When we subsequently applied the exact same methodology to a duration dimension, however, the results were in partial correspondence with both the decision-bound and the distribution theories. Subsequent model-based analyses confirmed the discrepancy between the two dimensions: For the formant frequency dimension, the decision-bound model gave the best quantitative account of the data, whereas the distribution model fitted the duration data best. We formulated NENA, a new, hybrid model of categorization that combines stochastic stimulus encoding with stochastic category activation.

The new model gave a better combined account of the data across the two stimulus dimensions than any of the other models, although some discrepancies between model and data persisted.

In our discussion of the experimental data, we asked why different categorization mechanisms seem to operate for the two stimulus dimensions. At present, we cannot provide an explanation for this difference. There is, however, a theoretical perspective that may guide future research into this matter. Stevens and Galanter (1957) introduced the concepts of *prothetic* and *metathetic* scales. A prothetic scale is a psychological scale to which, at a physiological level, an "additive" mechanism applies—that is, increasing a value on a prothetic scale is equivalent to adding more of the same. Examples of prothetic scales are brightness, loudness, and, in the present study, duration. A longer sound simply has "more duration" than a shorter sound and is presumably encoded at a physiological level by a stronger or longer firing of basically the same neurons. In contrast, a "substitutive" mechanism applies for metathetic scales, such as (visual) position, pitch, and, presumably, timbrelike magnitudes, such as formant frequency. A pure tone with a higher pitch does not simply have "more frequency" than one of a lower pitch. Instead, it essentially stimulates different fibers in the auditory nerve. Empirically, the difference between the two scales is evidenced by the fact that for metathetic scales the jnd measured in subjective units is constant across the scale (e.g., the jnd for pitch expressed in mels is the same for low and high tones), whereas the same does not hold for prothetic scales (the jnd for loudness expressed in sones is smaller at the low end of the scale than at the high end).

We hypothesize that either the storage of category representations or the comparison of a stimulus with a category is noisier for prothetic categories, such as duration, than for metathetic categories, such as formant frequency. According to this hypothesis, other prothetic auditory dimensions, such as loudness, should pattern with duration on a similar categorization task, whereas metathetic dimensions, such as pitch or dynamic timbre (formant transitions), should pattern with formant frequency. This is a topic for future research.

Finally, we return to the ultimate purpose of this research, which is to learn about the representations and processes underlying speech perception. As we mentioned in the introduction, four theories of phonetic categorization can be distinguished: decision-bound, prototype, distribution, and exemplar theories. Although they make fundamentally different claims about various aspects of categorization, it has proved extremely difficult to experimentally distinguish between the four alternatives. We know of only two studies in which researchers explicitly attempted to do so within the context of phonetic perception. Samuel (1982) contrasted the decision-bound and prototype accounts of phonetic categorization using selective adaptation in a /ga/–/ka/ categorization task. The experimental results showed that more adaptation was obtained for adaptors near the /ga/ prototype than for adaptors nearer to or farther away from the /ga/–/ka/ boundary. Samuel interpreted this as evidence in support of a prototype theory of phonetic categorization. The study was not conclusive, however. First, the evidence is in agreement not only with a prototype theory but also with distribution and exemplar theories, which Samuel did not explicitly test because they were less topical at the time. Furthermore, because the adaptation paradigm itself is not well understood and is the subject of dispute (see Remez, 1987), the evidence should be considered as relatively indirect.

A second study that explicitly contrasted theories of phonetic categorization was published by Nearey and Hogan (1986), who reanalyzed a set of production and perception data for the three-way voicing contrast in Thai stop consonants collected by Lisker and Abramson (1970). Lisker and Abramson measured voice-onset time on a set of naturally produced instances of the three voicing categories. Next, they constructed a synthetic stimulus continuum that varied in voice-onset time and asked listeners to categorize the stimuli according to voicing. Lisker and Ambramson noted the striking similarity of the crossover points in the production and perception data. Hoping to be able to use the data to distinguish between competing categorization theories, Nearey and Hogan (1986) fitted two formal categorization models to the data. The first was a decision-bound model that assumed noisy stimulus encoding and boundary locations that were (near) optimal given the production data. The second was a distribution-based model that assumed that the incoming stimulus was compared with the categories represented by the pdfs of the production data, followed by a choice based on the Luce (1963) choice rule. Both models fitted the data well, and the difference in goodness of fit was too small to warrant selection of one over the other. At present, it remains undecided which of the four categorization theories applies to phonetic categorization.

Although our study is intended as a first step toward solving this issue, we have to be cautious about generalizing our results to phonetic categorization. We can identify many differences between our stimuli and those in phoneme categorization. In particular, our stimuli consisted of nonspeech signals designed to resemble speech in crucial regards (duration and formant frequency). Recent research inspired by the current study sheds some light on these issues. Using training and testing protocols very similar to those in our study, Goudbeek, Smits, Cutler, & Swingley (2005) directly compared the acquisition of auditory and phonetic categories by adults. In the nonspeech condition, Dutch listeners categorized stimuli that simultaneously varied in duration and resonant frequency. In the speech condition, American listeners categorized stimuli that consisted of variants around the midpoints of three Dutch high-front vowels that do not occur in English and that differ in duration or frequency of the first formant. In terms of the speed of learning and the proportion of participants who eventually learned to do the task, the results for the nonspeech and speech stimuli were highly similar, indicating that findings based on nonspeech stimuli generalize to speech, at least within the experimental context used in these studies. This is of interest given the ongoing debate in the phonetic literature about the extent to which the perception of speech engages general auditory mechanisms or mechanisms that specifically evolved for the processing of speech (e.g., Liberman & Mattingly, 1985; Remez, Rubin, Berns, Pardo, & Lang, 1994). The finding that the current methodology obtains the same results for speech and nonspeech justifies its use as a valid means of studying phonemic categorization.

A final consideration for the use of multidimensional stimuli involves the acoustic variability that is a hallmark of natural speech. When researchers use stimuli that more closely mimic the degree of variability found in speech (e.g., variation in speaker, phonetic context, and speaking rate), the variability itself may affect the categorization mechanisms used. Recently, a number of exemplar models of speech perception have been proposed (Goldinger, 1997; Johnson, 1997a, 1997b). Rather than conceptualizing speech perception as a process wherein a more abstract representation is created from the myriad of highly variable and idiosyncratic tokens, some contemporary theories have emphasized the retention of detailed voice information in episodic representations. Goldinger and colleagues (Goldinger, 1997, 1998; Goldinger & Azura, 2004; Goldinger, Azura, Kleider, & Holmes, 2003) provided evidence for the existence of detailed episodic memory traces of spoken words in lexical access processes. An examination of speakers' recognition accuracy and listeners' imitation judgments shows sensitivity to previously encountered instances. Indexical aspects of speech are stored in memory and can be used later in perception and production. In a similar vein, Johnson (1997a) presented an exemplar-based model for vowel identification, taking into account aspects of talker variability that affect human vowel perception performance. Johnson used five acoustic parameters (F0, F1–F3, and vowel duration) as input to the model. The model's overall correct vowel identification was 80 human listeners' ability to identify vowels (Ryalls & Lieberman, 1982). Variability in speech that distinguishes speakers is retained in the set of exemplars. Both sets of data suggest that categorization takes place by reference to detailed auditory exemplars that preserve speaker-specific information, data most compatible with exemplar or distribution theories.

It is, at this point, difficult to predict to what extent the results of the present study generalize to the categorization of multidimensional speech sounds. This constitutes a topic for future research. We therefore merely use our results to formulate a number of hypotheses about phoneme categorization that researchers may test in future experiments using different stimuli and tasks.

Concerning the representation issue, the present results lead us to hypothesize that speech sounds are represented neither by prototypes nor by boundaries (rules) separating the speech sounds but instead by distributions. These distributions capture the natural variation of speech sounds, as encountered by the listener. Of course, because speech sounds are acoustically multidimensional, these distributions are multidimensional, too, in contrast to the unidimensional distributions of our experiments. On the basis of our results, we cannot decide whether the distributions are parametric—that is, economical summary descriptions—perhaps in the form of Gaussian pdfs, or nonparametric, perhaps in the form of multiple exemplars of previously heard sounds.

Concerning the processing issue, our results lead us to propose a new model combining aspects of the decision-bound and distribution models. In this model, the stimulus encoding is stochastic, as in the decision-bound model. Next, a similarity calculation is made, as in the distribution model, albeit stochastic. Finally, a deterministic choice is made, as in the decision-bound model. However, as the choice was based on a comparison of (noisy) activation levels, a decision-bound model, as such, did not play an explicit role in the choice process. We hypothesize that the phoneme categorization process has these same three components. Although the deterministic decision process is particular to the phoneme-categorization task and need not operate in the process of recognizing words or larger units, we hypothesize that the other two components are also active in the everyday recognition of running speech.

References

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

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

Ainsworth, W. A. (1968). Perception of stop consonants in synthetic CV syllables. *Language and Speech, 11,* 139–155.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716–723.

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association, 88,* 669–679.

Andruski, J. E., & Nearey, T. M. (1992). On the sufficiency of compound target specification of isolated vowels and vowels in /bVb/ syllables. *Journal of the Acoustical Society of America, 91,* 390–410.

Ashby, F. G., & Maddox, W. T. (1993). Relations 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., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review, 95,* 124–150.

Assman, P. F., Nearey, T. M., & Hogan, J. T. (1982). Vowel identification: Orthographic, perceptual and acoustic aspects. *Journal of the Acoustical Society of America, 71,* 975–987.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., & Gerstman, L. J. (1952). Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America, 24,* 597–606.

Durlach, N. I., & Braida, L. D. (1969). Intensity perception: I. Preliminary theory of intensity resolution. *Journal of the Acoustical Society of America, 46,* 372–383.

Ennis, D. M. (1988). Confusable and discriminable stimuli: Commentary on Nosofsky (1986) and Shepard (1986). *Journal of Experimental Psychology: General, 117,* 408–411.

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

Goldinger, S. (1997). Words and voices: Perception and production in an episodic lexicon. In K. Johnson & J. Mullennix (Eds.), *Talker variability in speech processing* (pp. 33–66). San Diego, CA: Academic Press.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105,* 251–279.

Goldinger, S., & Azura, T. (2004). Episodic memory reflected in printed word naming. *Psychonomic Bulletin & Review, 11,* 716–722.

Goldinger, S., Azura, T., Kleider, H., & Holmes, V. (2003). Font-specific memory: More than meets the eye? In J. Bowers & C. Marsolek (Eds.), *Rethinking implicit memory* (pp. 157–196). Oxford, England: Oxford University Press.

Goudbeek, M., Smits, R., Cutler, A., & Swingley, D. (2005). Acquiring auditory and phonetic categories. In H. Cohen & C. Lefebvre (Eds.), *Categorization in cognitive sciences* (pp. 497–513). Amsterdam: Elsevier.

Hillenbrand, J. M., & Nearey, T. M. (1999). Identification of resynthesized /hVd/ utterances: Effects of formant contour. *Journal of the Acoustical Society of America, 105,* 3509–3523.

Johnson, K. (1997a). The auditory/perceptual basis for speech segmentation. *OSU Working Papers in Linguistics, 50,* 101–113.

Johnson, K. (1997b). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. Mullennix (Eds.), *Talker variability in speech processing* (pp. 146–166). San Diego: Academic Press.

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

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics, 50,* 93–107.

Lee, W., & Zentall, T. R. (1966). Factorial effects in the categorization of externally distributed stimulus samples. *Perception & Psychophysics, 1,* 120–124.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology, 54,* 358–368.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1–36.

Lisker, L., & Abramson, A. (1970). The voicing dimension: Some experiments in comparative phonetics. In B. Hala, M. Romportl, & P. Janota (Eds.) *Proceedings of the 6th International Congress of Phonetic Sciences* (pp. 563–567). Prague, Czechoslavakia: Academia.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & S. E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 103–189). New York: Wiley.

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

Maddox, W. T., & Ashby, F. G. (1998). Selective attention and the formation of linear decision boundaries: Comment on McKinley and Nosofsky (1996). *Journal of Experimental Psychology: Human Perception and Performance, 24,* 301–321.

Massaro, D. W., & Cohen, M. M. (1983). Phonological context in speech perception. *Perception & Psychophysics, 34,* 338–348.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models.* London: Chapman & Hall.

Mermelstein, P. (1978). On the relationship between vowel and consonant identification when cued by the same acoustic information. *Perception & Psychophysics, 23,* 331–336.

Miller, J. L. (1994). On the internal structure of phonetic categories: A progress report. *Cognition, 50,* 271–285.

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

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.

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, FL: Academic Press.

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

Nosofsky, R. M. (1998). Selective attention and the formation of linear decision boundaries: Reply to Maddox and Ashby (1998). *Journal of Experimental Psychology: Human Perception and Performance, 24,* 322–339.

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 924–940.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review, 85,* 172–191.

Remez, R. E. (1987). Neural models of speech perception: A case history. In S. Harnad (Ed.), *Categorical perception* (pp. 199–225). Cambridge, England: Cambridge University Press.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review, 101,* 129–156.

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

Ryalls, J., & Lieberman, P. (1982). Fundamental frequency and vowel perception. *Journal of the Acoustical Society of America, 72,* 1631–1634.

Samuel, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics, 31,* 307–314.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 3–27.

Smits, R. (2001). Evidence for hierarchical categorization of coarticulated phonemes. *Journal of Experimental Psychology: Human Perception and Performance, 27,* 1145–1162.

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

Volaitis, L. E., & Miller, J. L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *Journal of the Acoustical Society of America, 92,* 723–735.

Whalen, D. H. (1989). Vowel and consonant judgments are not independent when cued by the same information. *Perception & Psychophysics, 46,* 284–292.

Appendix

Derivation of Predictions of Basic Models

Prototype Theory

If we assume Gaussian similarity functions (e.g., Nosofsky, 1986), the probability $p(A|S_i)$ of assigning stimulus S_i, defined by parameter value ψ_i, to category A is given by

$$p(A|S_i) = \frac{\eta_A(\psi_i)}{\eta_A(\psi_i) + \eta_B(\psi_i)}$$

$$= \frac{\exp - k(\psi_i - \mu_A)^2}{\exp - k(\psi_i - \mu_A)^2 + \exp - k(\psi_i - \mu_B)^2}$$

$$= \frac{1}{1 + \exp - 2k(\mu_A - \mu_B)[\psi_i - 1/2\,(\mu_A + \mu_B)]}, \qquad \text{(A1)}$$

where $\eta_A(\psi_i)$ is the similarity of ψ_i to category A and k is a sensitivity parameter (e.g., Ashby & Maddox, 1993). Of course, $p(B|S_i) = 1 - p(A|S_i)$. $p(A|S_i)$ is a logistic function of ψ_i. The function's inflection point, which corresponds to the value of ψ_i where $p(A|S_i) = ½$, is located at $\psi_i = ½(\mu_A + \mu_B)$—that is, halfway between the two means. The slope s of the logistic function, defined as the absolute value of the coefficient of ψ_i in the exponent of Equation A1, equals $2k(\mu_A - \mu_B)$. Thus, the slope of the categorization function is proportional to the distance between the means of the pdfs used in the training phase. For the purpose of Figure 1, we set k to 0.25; we chose this value such that the categorization functions were comparable to those for the other theories.

Distribution Theory

The similarity $\eta_A(\psi_i)$ of a given stimulus S_i to category A is assumed to be equal to the "unnormalized" likelihood $p(\psi_i|A)$ that the stimulus was produced by the particular category:

$$\eta_A(\psi_i) = p(\psi|A) \cdot \sigma\sqrt{2\pi}$$

$$= \exp - \frac{k}{2\sigma^2}(\psi_i - \mu_A)^2, \qquad \text{(A2)}$$

where k is a sensitivity parameter. The term *unnormalized* refers to the assumption that self-similarity equals unity—that is, $\eta_A(\mu_A) = 1$—which has the effect that the similarity function is generally not a probability density function because its integral differs from one.

Finally, we assume that response probabilities are calculated from similarity functions via Luce's (1963) choice rule. This leads to the following expression for $p(A|S_i)$:

$$p(A|S_i) = \frac{\eta_A(\psi_i)}{\eta_A(\psi_i) + \eta_B(\psi_i)}$$

$$= \frac{\exp - \frac{k}{2\sigma^2}(\psi_i - \mu_A)^2}{\exp - \frac{k}{2\sigma^2}(\psi_i - \mu_A)^2 + \exp - \frac{k}{2\sigma^2}(\psi_i - \mu_B)^2}$$

$$= \frac{1}{1 + \exp - \frac{k(\mu_A - \mu_B)}{\sigma^2}[\psi_i - 1/2\,(\mu_A + \mu_B)]}. \qquad \text{(A3)}$$

As was the case for the prototype theory, $p(A|S_i)$ is a logistic function of ψ_i with inflection point at $½(\mu_A + \mu_B)$. The categorization function's slope s is now equal to $\frac{k}{\sigma^2}(\mu_A - \mu_B)$—that is, it is proportional to the distance between the means of the training pdfs divided by their variance. For the purpose of Figure 1, we set the value of k to 1.

Discrimination of Test Continua

Method

Participants. We recruited 8 students at Nijmegen University as participants. All reported normal hearing and had Dutch as their native language. None participated in the categorization experiments reported in the article.

Stimuli. We used stimulus numbers 1, 3, 5, 7, 9, and 11 of the formant frequency and duration continua. We did not use all stimuli to limit the size of the experiment.

Procedure. We adopted a same–different (AX) paradigm. On a given trial, we played two stimuli after each other, with an interstimulus interval of 300 ms. The two stimuli were either the same or different, in which case they were two steps apart on the stimulus continuum (e.g., Stimuli 3 and 5). Participants were seated in a soundproof booth in front of a computer screen. Stimuli were presented binaurally through Sennheiser headphones. After hearing a pair of stimuli, participants were required to indicate whether they thought the stimuli were the same or different by pressing one of two appropriately labeled buttons. After the button press, the correct answer was displayed briefly on the screen, and a new trial was initiated.

The experiment consisted of two parts: a subexperiment testing formant frequency discrimination and one testing duration discrimination. Half the participants started with formant frequency discrimination; the other half started with duration discrimination. After 20 practice trials, participants were presented with five blocks of 40 trials each. Every "different" pair was presented four times in each block, twice in ascending order (e.g., 3–5) and twice in descending order (5–3). The "same" pairs (e.g., 3–3) were each presented four times per block, except for the pairs 1–1 and 11–11, which were presented twice per block. Thus, the probabilities of being presented with a same or different pair were equal. Within blocks, stimuli were pseudorandomized, with different randomizations for different participants. After the five experimental blocks, participants had a short break, in which we explained to them that they were to do the experiment again, but this time the sounds would differ from each other in another way. They then started on the second subexperiment, in which the stimuli varied along the other stimulus dimension. The procedure of the second subexperiment was identical to that of the first, including the practice trials.

Results and Discussion

For unknown reasons, 1 of the participants performed below chance level on the duration stimuli and above chance but still worse than the other 7 participants for the formant frequency stimuli. We removed this participant's data. Using Macmillan and Creelman's (1991) Table A5.4 (differencing model), we calculated d' values for each stimulus pair for each participant. Figure A1 presents means and standard deviations of d' for the two stimulus continua.

(*Appendix continues*)

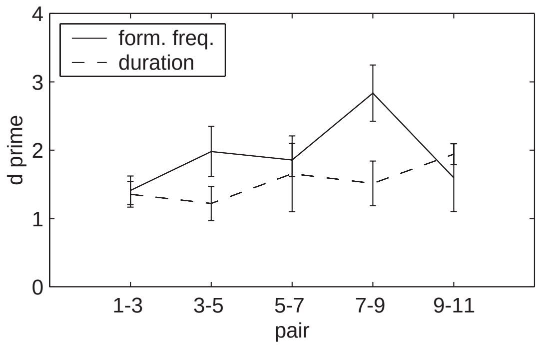

Figure A1. Discriminability, expressed as d', as a function of stimulus pair for the formant frequency (form. freq.) and duration continua. Error bars represent plus or minus one standard error.

The results of the discrimination experiment tell us, first, that the stimuli of both continua were moderately confusable. Average d's for a two-step distance along the formant-frequency and duration continua were 1.9 and 1.5, respectively. If we assume that d's are additive along a one-dimensional continuum, average d's for the discrimination of two consecutive stimuli are 1.0 and 0.8 for the formant frequency and duration continua, respectively (i.e., two consecutive stimuli are on the border of being discriminable). Average d's for the discrimination of the endpoint stimuli were 9.7 and 7.7, respectively (i.e., endpoint stimuli were highly discriminable).

We ran an ANOVA with d' as the dependent variable, pair number and stimulus dimension as fixed factors, and participant as random factor. Stimulus dimension did not prove significant, $F(1, 6) = 3.3$, $MSE = 0.84$, which means that average d's were the same for the formant frequency and duration continua. Pair number did not reach significance either, $F(4, 24) = 1.9$, $MSE = 0.62$, which means that d' was constant across pairs. Of the possible interaction terms, only Pair Number $\times$ Stimulus Dimension reached significance, $F(4, 24) = 3.2$, $p < .05$, $MSE = 0.461$, $\eta^2 = .35$. Individual t tests for the difference in discriminability between formant frequency and duration for each of the stimulus pairs showed that the interaction was due to Stimulus Pair 7–9, which was the only pair for which d' was significantly different for the two dimensions, $t(6) = -3.6$, $p < .02$.

In addition to the ANOVA, we ran separate linear regressions for the formant frequency and duration data, with d' as the dependent variable and pair number as the independent variable. For neither dimension did the pair number factor reach significance. This shows that there is no linear trend in d', giving further support for the constancy of discriminability of both continua.

From the discrimination experiments, we draw the following conclusions. First, the two stimulus continua we used in our categorization experiments were of equal discriminability. Both covered the same number of jnds. Second, consecutive stimuli were confusable, whereas continuum endpoints were highly discriminable. Finally, discriminability was almost constant across both continua, which means that stimuli on both continua were almost equidistant in the perceptual sense. The only exception to this rule is Stimulus Pair 7–9 in the formant frequency series, which was slightly more discriminable than the other pairs.