Journal of Experimental Psychology:
Human Perception and Performance
2001, Vol. 27, No. 5, 1145–1162

0096-1523/01/$5.00 DOI: 10.1037//0096-1523.27.5.1145

Evidence for Hierarchical Categorization of Coarticulated Phonemes

Roel Smits
Max Planck Institute for Psycholinguistics

The reported research investigates how listeners recognize coarticulated phonemes. First, 2 data sets from experiments on the recognition of coarticulated phonemes published by D. H. Whalen (1989) are reanalyzed. The analyses indicate that listeners used categorization strategies involving a hierarchical dependency. Two new experiments are reported investigating the production and perception of fricative–vowel syllables. On the basis of measurements of acoustic cues on a large set of natural utterances, it was predicted that listeners would use categorization strategies involving a dependency of the fricative categorization on the perceived vowel. The predictions were tested in a perception experiment using a 2-dimensional synthetic fricative–vowel continuum. Model analyses of the results pooled across listeners confirmed the predictions. Individual analyses revealed some variability in the categorization dependencies used by different participants.

How do listeners recognize coarticulated speech? The acoustic realizations of linguistic units like phonemes or syllables are, because of coarticulation, considerably affected by neighboring units in the speech stream. As a result, listeners have to deal with great variability in the acoustic representation of linguistic units. How listeners solve this problem remains one of the central issues in research on speech perception.

Hypotheses on how listeners process acoustic information to recognize linguistic units all have three essential components. First, there is the issue of the recognition unit. Assuming that coarticulation spreads out over a given number of phonemes, or a given time interval, the acoustic realization of a large linguistic unit, such as the syllable or word, will logically be less variable than that of a small unit, such as the phoneme. This argument suggests that using a large unit would simplify the recognition problem, but it comes at a cost. With increasing unit size, the sublexical repertoire quickly expands, which in turn increases the demands on memory and processing when recognizing these units.

The second issue concerns the use of acoustic context in recognition. What is the nature of the time windows across which information is used for the recognition of successive linguistic units? Concerning this issue, Fowler (1984) has distinguished two basic hypotheses. According to the first, the speech signal is segmented into discrete, nonoverlapping portions, each of which correspond to a single linguistic symbol. In the second hypothesis, successive symbols use information from relatively wide, overlapping windows, thus sharing acoustic information between successive symbols.

Finally, there is the issue of phonological context. Given that the acoustic realization of a linguistic symbol depends on the surrounding symbols, it may be advantageous to make the processing of the acoustic information dependent on the surrounding symbols. If, for example, the acoustic realization of a fricative is very different when it is followed by a rounded versus unrounded vowel, it may be useful to the listener to make the fricative recognition dependent on whether or not the following vowel is perceived to be rounded.

These three components of the problem of the recognition of coarticulated speech—that is, the size of the recognition unit and the use of acoustic and phonological context—are, although theoretically distinct, obviously related. All three suggest opportunities for simplifying the problem by integrating more context information into the recognition. The first does so by making the recognition unit larger, the second by using a wider stretch of the acoustic signal, and the third—choosing a sort of middle way—by making the recognition of successive symbols interdependent. Many experiments have shown that phonetic context plays a role in speech perception (e.g., Mann, 1980; Mann & Repp, 1980; Schatz, 1954; Whalen, 1989), but few have convincingly demonstrated from which of the three components the observed context effects derive.

In recent years a pattern-classification framework has been formulated (Massaro, 1998; Nearey, 1990, 1997; Smits, 1997), which allows for a systematic study of the issue of how listeners recognize coarticulated speech, while keeping the three components to the recognition process apart. The framework views the issue as a pattern-classification problem, and listeners are assumed to essentially behave like multidimensional pattern recognizers.

For two decades, Massaro and colleagues (e.g., Massaro, 1998; Massaro & Oden, 1980) have supported a model of speech perception based on the fuzzy-logical model of perception (FLMP). This model assumes (a) syllable-sized recognition units, (b) nonoverlapping acoustic windows for successive phonemes, and (c) no

Parts of this research were carried out at the Department of Phonetics and Linguistics, University College London, United Kingdom. These parts were supported by North Atlantic Treaty Organization Science fellowship S 30-440 and Training and Mobility of Researchers fellowship ERBFMBICT950313 granted by the European Commission.

I am grateful to Pieter Meima for help in carrying out the experiments; to Stuart Rosen, Louis ten Bosch, Terry Nearey, Anne Cutler, and James McQueen for advice and encouragement; to Robert Remez, Doug Whalen, Jose Benki, and Gary Weismer for comments on an earlier version of this article; and to Doug Whalen (again) for permission to use his data sets.

Correspondence concerning this article should be addressed to Roel Smits, Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, the Netherlands. Electronic mail may be sent to Roel.Smits@mpi.nl.

phonological context effects. Whalen (1989), Batchelder and Crowther (1997), and Smits (in press) described how the assumption of nonoverlapping acoustic windows, though it has never been explicitly discussed by Massaro and colleagues, follows from the structure of the model when it is applied to four-alternative forced-choice data. This makes FLMP unsuitable as a model of the processes studied in the present research because previous analyses by Whalen (1989) and Nearey (1990) have shown that, in certain cases, acoustic cues are shared between categorizations of successive phonemes.

Nearey (1990) formulated a model for the analysis of phonetic categorization data on the basis of logistic regression (LR). Applications of the LR model to his own experimental data and those of others has led Nearey (1990, 1997, in press) to support a model of speech perception that has the following characteristics: (a) phoneme-sized recognition units, (b) overlapping acoustic windows for successive phonemes, and (c) no phonological context effects.

Although Nearey's approach has been very successful, and indeed inspired much of the present research, it has not been spared criticism. First, the interpretation of the *diphone bias* term in the LR model remains uncertain (Nearey, in press; Whalen, 1992). Second, although Nearey makes a general, qualitative link between the structure of his model and statistical distributions of acoustic cues in speech, the link is never tested explicitly (Smits, in press). Finally, Nearey's analysis framework does not allow for testing for specific, economical forms of phonological dependencies. If, in terms of his LR model, stimulus-tuned diphone terms were found to make a significant contribution, he would reject the phoneme as the recognition unit in favor of a larger unit, although this evidence might be accounted for by a phoneme recognizer that allows for phonological dependencies (Smits, in press).

A Theory of Hierarchical Phoneme Categorization

Smits (in press) presented a theory of hierarchical categorization of coarticulated phonemes. Like Nearey's theory, it assumes phoneme-sized recognition units and overlapping acoustic windows for successive phonemes, but in contrast with Nearey's theory, phonological context effects are allowed. The theory incorporates a very specific type of phonological context effects, namely, hierarchical context effects, in which one categorization depends on the other, but not vice versa. Moreover, Smits (in press) described why and how such hierarchical categorizations may be beneficial to listeners on the basis of an analysis of the acoustic consequences of coarticulation, thus trying to fill in the theoretical gaps mentioned above.

The article started with an analysis of the possible effects of coarticulation on statistical distributions of acoustic cues. The analysis focused on a case of strong coarticulation as occurs in the Dutch fricative–vowel syllables /si sy ʃi ʃy/. In these syllables, rounding spreads from the vowel to the preceding fricative, which affects the fricative spectrum.

Assume that both the fricative and the vowel contrast each have one dominant acoustic correlate: a fricative resonance F_{fr} for the /s/–/ʃ/ contrast and the third formant frequency $F3$ for the /i/–/y/ contrast (the choices of these acoustic dimensions are substantiated in a later section). In the hypothetical case characterized by the absence of coarticulation, F_{fr} would not be affected by the vowel, nor would $F3$ be affected by the fricative. In this case, the means of the probability-density functions (PDFs) for F_{fr} and $F3$ for the four syllables would form a rectangular arrangement in the $F_{fr} \times F3$ space. Figure 1A gives a visual impression of this case, where the circles represent isoprobability contours of the PDFs. An independent phoneme recognizer, using a horizontal vowel boundary and a vertical fricative boundary, will perform optimally.

Of course, in real life the fricative–vowel syllables are produced with coarticulation. Smits (in press) discussed two distinct effects of coarticulation on the arrangement of the PDFs. The first is the shift pattern shown in Figure 1B. Because lip rounding is already effective during the production of the fricative, the fricative resonance F_{fr} is lowered in frequency. If the frequency shift were equal for /s/ and /ʃ/, the pattern of Figure 1B would arise.

The distributional pattern of Figure 1B calls for a classification strategy that is different from that for Figure 1A. Several strategies are possible. First, a slanted fricative boundary can be used. This means that the fricative classification uses both F_{fr} and $F3$, which corresponds to using a wider acoustic window for the fricative classification, that is, one that overlaps with the window for the vowel classification. Alternatively, the classifier can make the classification of one contrast dependent on the other. For example, the position of the /s/–/ʃ/ boundary can be adjusted depending on the perceived vowel.

The second possible effect of coarticulation on the arrangement of the syllable PDFs is the convergence pattern of Figure 1C. Because of rounding in the /y/ context, the values of F_{fr} for /s/ and /ʃ/ might become more similar. Smits (in press) discussed why, in this case, it would be advantageous for a pattern classifier to be more fuzzy on the fricative classification in the /y/ context than in the /i/ context.

In general, rounding assimilation, and perhaps other forms of coarticulation, too, are expected to produce a combination of the shift and convergence patterns of Figures 1B and 1C, as illustrated in Figure 1D. The following crude but adequate acoustic–phonetic analysis of the production of fricative–vowel syllables supports this claim. If the length of the front cavity for fricatives /s/ and /ʃ/

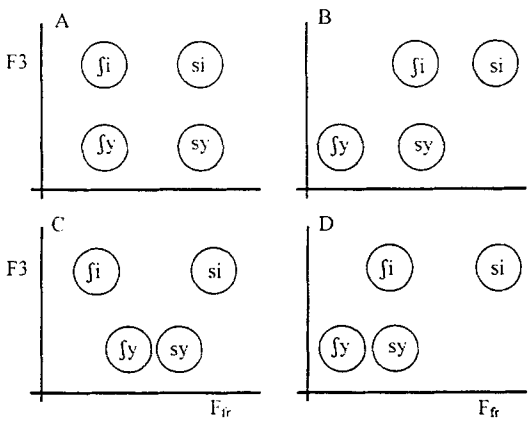

Figure 1. Geometries of acoustic cue distributions for four hypothetical varieties of coarticulation. Circles represent isoprobability contours of two-dimensional Gaussian probability-density functions. A: Absence of coarticulation; B and C: Shifted and converged geometries, respectively, due to coarticulation; D: A combination of B and C.

is estimated at 2 cm and 3 cm, natural resonances are predicted to occur at frequencies F_{fr} of 4.3 kHz and 2.8 kHz, respectively. Assuming that lip rounding adds a fixed amount of 0.5 cm to this front cavity, one would expect resonance frequencies of 3.4 kHz and 2.4 kHz for /s/ and /ʃ/, respectively, in vowel context /y/. Thus, the finding is that lip rounding causes F_{fr} to shift down both for /s/ and for /ʃ/ but by a larger amount for /s/ (0.9 kHz) than for /ʃ/ (0.4 kHz). This crude calculation shows that at least in the case of rounding assimilation, a combination of the shift and the convergence pattern can be expected.

Smits (in press) discussed how a pattern classifier, and presumably a listener, could deal with the various distributional patterns by using hierarchical categorizations, that is, categorizations in which one of the categorizations depends on the other, but not vice versa. Three types of hierarchical dependencies were distinguished: dependence of the position, orientation, or steepness of one phoneme boundary on the perceived value of the other phoneme.

Smits (in press) discussed two possible information-processing architectures implementing such hierarchical dependencies, one serial and the other parallel. In a serial architecture, one of the categorizations would have to wait for the other to finish. This scenario was seen as unrealistic for several reasons (see Smits, in press). A parallel architecture was instead proposed in which independent fuzzy categorizations were made initially, after which the (fuzzy) output of one of the categorizations was adjusted on the basis of the (fuzzy) output of the other.

Subsequently, the HICAT model of Hierarchical Categorization of coarticulated phonemes was introduced, which contains the three types of dependencies mentioned above. HICAT can be used to test for hierarchical dependencies in categorization by fitting the model to experimental categorization data and carrying out statistical tests on the significance of various model parameters. The model incorporates the following assumptions:

1. There exists a relatively simple monotonic mapping between physical stimulus dimensions and associated psychological dimensions. For example, in the case of auditory frequencies, this mapping is approximated by the equivalent rectangular bandwidth (ERB) scale (Glasberg & Moore, 1990).

2. Categorization probabilities for each contrast are linear logistic functions of the psychological dimensions.

3. One of the categorizations depends on the outcome of the other, but not vice versa. For example, in the fricative–vowel case the fricative categorization might depend on the perceived vowel.

Below, I give a short mathematical description of the HICAT model. For a full definition, see Smits (in press). The model is defined in terms of the recognition of syllables /si sy ʃ i ʃ y/, as a function of acoustic dimensions F_{fr} and $F3$ but can of course be applied to any four-alternative forced-choice classification involving two binary distinctions and two physical dimensions. The HICAT model assumes that one of the binary categorizations is made independently of the other. Suppose that this assumption holds for the vowel classification. HICAT assumes that the probabilities of classifying stimulus S_i as /i/ or /y/ depend on a variable α through a logistic function:

$$\log \frac{p(/\mathrm{i}/|S_i)}{p(/\mathrm{y}/|S_i)} = \alpha, \tag{1}$$

where $p(L|S_i)$ indicates the probability of assigning label L to stimulus S_i. The α is a linear combination of two psychological dimensions $\tilde{F}_{fr}$ and $\tilde{F}3$, which are the frequencies of the fricative resonance F_{fr} and vowel formant $F3$ coded in terms of ERB rate:

$$\alpha = p_0 + p_1\tilde{F}_{fr} + p_2\tilde{F}3. \tag{2}$$

The fricative categorization is made dependent on the vowel categorization. This means that the probabilities of classifying a stimulus as /s/ or /ʃ/ are conditional on the perceived vowel:

$$\log \frac{p(/\mathrm{s}/|/\mathrm{i}/, S_i)}{p(/\mathrm{ʃ}/|/\mathrm{i}/, S_i)} = \beta + (c_0 + c_\alpha\alpha + c_\beta\beta) \quad \text{and} \tag{3}$$

$$\log \frac{p(/\mathrm{s}/|/\mathrm{y}/, S_i)}{p(/\mathrm{ʃ}/|/\mathrm{y}/, S_i)} = \beta - (c_0 + c_\alpha\alpha + c_\beta\beta). \tag{4}$$

The parenthetical term in the right-hand terms of Equations 3 and 4 represents the dependency of the fricative categorization on the vowel categorization. β, like α, is a linear function of psychological dimensions $\tilde{F}_{fr}$ and $\tilde{F}3$:

$$\beta = q_0 + q_1\tilde{F}_{fr} + q_2\tilde{F}3. \tag{5}$$

The categorization dependency defined by the parenthetical term of Equations 3 and 4 has three components, associated with parameters c_0, c_α, and c_β:

c_0: Dependency of the *position* of the /s/–/ʃ/ boundary on the perceived vowel;

c_α: Dependency of the *orientation* of the /s/–/ʃ/ boundary on the perceived vowel;

c_β: Dependency of the *steepness* of the /s/–/ʃ/ boundary on the perceived vowel.

The full HICAT model defined above has nine free parameters: stimulus coefficients p_0, p_1, p_2, q_0, q_1, q_2 and dependency parameters c_0, c_α, c_β. Nested under the full model is the independent categorization model, for which $c_0 = c_\alpha = c_\beta = 0$. This model captures the situation where the two phonemes are perceived in a statistically independent fashion, that is,

$$p(/\mathrm{si}/) = p(/\mathrm{s}/)p(/\mathrm{i}/), \tag{6}$$

and similarly for the other syllables. The independent model coincides with Nearey's four-alternative LR model without any diphone terms.

The HICAT model can be used to analyze phoneme categorization data for the presence of hierarchical categorization dependencies. Smits (in press) also introduced a method for predicting categorization dependencies from measurements of acoustic cues on natural utterances. Like the HICAT model, this prediction technique was based on the theoretical analysis of the influence of coarticulation on the distributions of acoustic cues. Using the technique, predictions can be made about the direction and type of hierarchical dependencies most likely to be used by listeners, assuming they behave like statistical pattern recognizers.

In the present article, I applied the theory and methods introduced in Smits (in press) to experimental data. First, I analyzed two existing data sets by Whalen (1989) using HICAT. Whalen (1989) carried out two experiments that are particularly relevant to the problem of the recognition of coarticulated phonemes. In the

first experiment, listeners were asked to classify vowel height and final stop voicing in synthetic consonant–vowel–consonant (CVC) words, using the labels *bad, bat, bed,* and *bet.* In the synthetic stimuli, *F*1 and vocoid duration were varied orthogonally. As both vowel height and final stop voicing are known to be strongly affected by vocoid duration, it was interesting to investigate how listeners would categorize these stimuli. In a different experiment, Whalen varied the frequency of a fricative pole and *F*2 of the following vowel in synthetic fricative–vowel stimuli. Listeners were asked to classify the stimuli as *see, Sue, she,* or *shoe.* This experiment investigated the situation where listeners have to deal with strong coarticulatory effects on a relatively stationary portion of the acoustic signal. Because of spreading of the feature (round) from the vowel to the preceding fricative, the fricative spectrum is highly affected by the vowel. Whalen (1989) concluded from his experiments that coarticulated phonemes are recognized in a mutually dependent fashion. An LR analysis of these data performed by Nearey (1990), on the other hand, led to the conclusion of independent phoneme recognizers. The present study provides yet a different interpretation of Whalen's data involving hierarchical dependencies.

A problem with the fricative–vowel experiment of Whalen (1989) was that only very few stimuli were used. This problem made the conclusions of the HICAT analyses of these data rather unreliable. Because the fricative–vowel case is so relevant to the issue under investigation, I decided to subject it to more detailed study. Besides, Whalen's study did not provide acoustic data to which the results of the perception experiments could be compared. Therefore, two new experiments were carried out investigating the production and perception of fricative–vowel syllables in Dutch. In Experiment 1, a large set of tokens of the Dutch syllables /si sy ʃi ʃy/, uttered by many speakers, were collected. The frequency of the main fricative resonance and *F*3 in the vowel were measured on these utterances. On the basis of the acoustic measurements, predictions were made of the direction and type of categorization dependency that listeners were most likely to use in their categorization of the syllables /si sy ʃi ʃy/. These predictions were tested in Experiment 2. I constructed a synthetic, two-dimensional, fricative–vowel continuum by varying the frequency of the fricative resonance and vowel *F*3 orthogonally. These stimuli were presented to listeners for categorization as /si/, /sy/, /ʃi/, or /ʃy/. The HICAT model was applied to the resulting data to test for dependency direction and types.

The remainder of the article is structured as follows. The next section describes the experiments reported by Whalen (1989), the reanalyses of the data by Nearey (1990), and the HICAT reanalyses. Next, the production experiment is presented and predictions are made of the categorization strategies used by listeners. In the subsequent section, the perception experiment is presented, followed by the HICAT model analyses. The article concludes with a summary and discussion.

Reanalyses of Whalen (1989)

Mermelstein (1978) carried out two perception experiments that tested for processing dependencies in the categorization of successive phonemes. Mermelstein varied steady state *F*1 and vocoid duration in synthetic CVC stimuli and asked listeners to classify the stimuli as either *bad, bat, bed,* or *bet.* Henceforth, this experiment is indicated as the *bad bet* experiment. As vocoid duration is known to influence the recognition of both vowel and final consonant voicing in English, it was interesting to investigate the processing dependencies in this case: Do vowel and consonant categorizations simply share acoustic information while remaining independent otherwise, or are there dependencies of the type where one categorization depends on the outcome of the other? On the basis of his experimental data, Mermelstein favored the former conclusion: Both vowel and consonant categorizations use the duration cue, but vowel categorization does not depend on the perceived consonant, nor vice versa.

Whalen (1989) conducted a replication of Mermelstein's *bad bet* experiment plus two additional ones investigating the dependencies in perception of fricative–vowel syllables. Compared with Mermelstein's Experiment 1, Whalen collected eight times as many responses and carried out somewhat different analyses. Contrary to Mermelstein, Whalen found dependencies of the location of the stop voicing boundary on the perceived vowel and of the vowel boundary on perceived stop voicing.

In Experiment 3 (henceforth, the *sushi* experiment), Whalen varied both the frequency of a fricative pole and vowel *F*2 in synthetic fricative–vowel stimuli and asked listeners to classify the stimuli as either *see, Sue, she,* or *shoe.* Production of fricative–vowel syllables is subject to regressive assimilation: If the vowel is rounded, rounding spreads to the preceding fricative, lowering fricative resonances in frequency. Experiments by Whalen (1981) and Mann and Repp (1980) had already shown that listeners compensate for this coarticulatory effect by accordingly shifting the fricative boundary with the following vowel. The results of Whalen's *sushi* experiment replicated this effect: The fricative boundary location depended on the perceived vowel, and vice versa, analogously to the *bad bet* case.

Nearey (1990) reanalyzed Experiments 1 (*bad bet*) and 3 (*sushi*) of Whalen (1989) using LR. The advantage of the LR analysis over Whalen's analyses is that a complete two-dimensional data set can be analyzed in a single model, allowing for explicit statistical tests of the processing dependencies of interest. In addition, territorial plots of selected models provide easy-to-interpret illustrations of the various inferred processing dependencies. Nearey's analysis of Whalen's *bad bet* experiment showed that first of all, vowel as well as consonant categorization depend on both acoustic cues. In addition, like Whalen, Nearey found a dependency of the consonant boundary location on the vowel and vice versa. Nearey's reanalysis of the *sushi* data produced the same pattern, except that the vowel categorization was found to be independent of the frequency of the fricative pole.

Nearey's (1990) interpretation of these findings was, however, fundamentally different from Whalen's (1989). Within the LR framework, the mutual dependency of the boundary location is captured by a single bias parameter. This is compatible with a processing architecture in which the phoneme categorizations are independent, and a diphone bias is effective only in the decision stage. Nearey (1997) replicated the *bad bet* experiment on a larger scale and found the same basic pattern: independent vowel and final consonant classifications plus diphone bias effects. On the basis of these results, Nearey (1997) favors a model based on independent phoneme categorizations (which may share acoustic information), whereas Whalen supports a model in which there is information exchange between the phoneme categorizations.

Below, further reanalyses of Whalen's *bad bet* and *sushi* data are given using the HICAT model. Nearey's LR model gives a good and very parsimonious account of Whalen's data sets. However, the LR model does not incorporate hierarchical dependencies and therefore cannot test for them. HICAT explicitly models such dependencies, so it can be used to test whether such hierarchical dependencies could give as good an account of the data.

Bad Bet Experiment

Whalen's (1989) Experiment 1 was an almost exact replication of Mermelstein's (1978) Experiment 1. Whalen created a 3 × 10 synthetic CVC continuum by varying vocoid duration D_v in 10 steps and *F*1 in the steady state of the vowel in three steps. Listeners were required to label each stimulus as either *bad, bat, bed,* or *bet.* Each of 16 listeners classified each stimulus 20 times. Results were pooled across listeners, yielding 320 judgments per stimulus in total. For further details, see Whalen (1989).

In accordance with the procedure proposed in Smits (in press), HICAT model analyses started with three model fits. First, the independent model (henceforth indicated as $V - C$) was fitted; that is, dependency parameters c_0, c_α, c_β were set to zero, leaving six free parameters. In this case, the model is equal for the two dependency directions (because there is no dependency) and is also equal to Nearey's independent model without the diphone bias (Nearey, 1990, Model 2, p. 359). Following Nearey (1990), it was assumed that the psychological representation of vocoid duration D_v was equal to the square root of the vocoid duration $\sqrt{D_v}$ (see also Abel, 1972). Next, the full dependency model, including all three dependency parameters, was fitted separately for the two dependency directions. Recall that dependency parameters c_0, c_α, and c_β code for dependency of boundary position, orientation, and steepness, respectively, of one categorization on the outcome of the other categorization. One of the two dependency directions, henceforth indicated as $V \rightarrow C$, assumed that the final consonant recognition depended on the vowel; the other ($V \leftarrow C$) assumed the reverse dependency. Both models had nine free parameters. The modeling results are given in the top three rows of Table 1.

Table 1 presents results in similar fashion to Nearey's (1990) Table 4. G^2, also known as *deviance,* is a goodness-of-fit (GOF) measure appropriate for the analysis of categorical data (e.g., Agresti, 1990), with smaller deviance indicating better fit. In all model fits, parameter values were found that minimized G^2. A popular alternative GOF measure is the root mean squared difference (RMSD) between expected and observed probabilities (see also Massaro & Oden, 1980; Nearey, 1990). Although RMSD is an intuitively more appealing GOF measure, it is statistically less appropriate for categorical data, and here it is only used to give additional feel for the GOF levels obtained in the model analyses. Because different researchers seem to calculate RMSD in slightly different ways, the definition of RMSD as it was calculated for the present research is given in Equation 7.

$$\text{RMSD} = \sqrt{\left(\frac{1}{4N}\sum_i \{[p_e(\text{bad}|S_i) - p_o(\text{bad}|S_i)]^2 + [p_e(\text{bat}|S_i) - p_o(\text{bat}|S_i)]^2 + [p_e(\text{bed}|S_i) - p_o(\text{bed}|S_i)]^2 + [p_e(\text{bet}|S_i) - p_o(\text{bet}|S_i)]^2\}\right)}, \quad (7)$$

Table 1
Goodness-of-Fit Analysis for Various Models for the Categorization Data of the Bad Bet Experiment by Whalen (1989)

Model	Parameters	G^2	RMSD (%)	*df* res.	Overdisp.
1	$V - C$, full	226	3.28	84	2.69
2	$V \rightarrow C$, full	139	2.84	81	1.72
3	$V \leftarrow C$, full	161	2.82	81	1.99
4	$V \rightarrow C$, c_0	161	2.88	83	1.94
5	$V \rightarrow C$, c_α	188	3.27	83	2.27
6	$V \rightarrow C$, c_β	191	3.24	83	2.30
7	$V \rightarrow C$, $c_0 + c_\alpha$	145	2.85	82	1.77
8	$V \rightarrow C$, $c_0 + c_\beta$	144	2.86	82	1.73
9	LR + diph. bias	165	2.92	83	1.99

Note. $V - C$ refers to the independent model, containing 6 parameters p_0, p_x, p_y, q_0, q_x, q_y; $V \rightarrow C$ refers to the model in which the final stop categorization depends on the vowel; $V \leftarrow C$ refers to the model with the opposite hierarchy. c_0 indicates that c_0 was the only dependency parameter included in the model. LR + diph. bias refers to Nearey's (1990) logistic regression model that includes a diphone bias but no stimulus-tuned diphone terms. RMSD = root mean square difference; *df* res. = residual degrees of freedom; overdisp. = overdispersion.

where N is the number of stimuli, the summation is made across all stimuli S_i, and $p_e(\text{bad}|S_i)$ and $p_o(\text{bad}|S_i)$ indicate expected and observed probabilities of response *bad* to stimulus S_i. Overdispersion (McCullagh & Nelder, 1989; Nearey, 1990) equals G^2 divided by the number of degrees of freedom (*df*s) of the residual and is used in quasi-likelihood significance tests below.

Table 1 shows that compared with the independent model, G^2 is strongly reduced by allowing for hierarchical dependencies in the form of the three dependency parameters c_0, c_α, c_β. Both dependency directions give large improvements over independence, but $V \rightarrow C$, with the consonant recognition depending on the vowel, produces a better fit than $V \leftarrow C$, where the vowel recognition depends on the consonant.

The finding that both dependency directions give improvements over the independent model should not be interpreted as evidence for the presence of both dependencies in listeners' categorizations. Technically speaking, the independence model is nested under the full dependency model for both directions. In consequence, the dependency models for both directions always fit the data at least as well as the independence model. Moreover, the earlier theoretical analysis explained at a more intuitive level how both categorization hierarchies allow for improved phoneme recognition, relative to the independent case, for the two hypothesized consequences of coarticulation on the acoustic cue distributions. On the basis of Monte Carlo simulations, Smits (in press) showed that under certain conditions a truly hierarchical categorizer may produce categorization data on the basis of which it is very difficult to decide what the direction of the categorization dependency actually was. Therefore, Smits provided a technique for estimating the reliability of choosing the best-fitting dependency direction as the correct direction. For the present data, the probability that $V \rightarrow C$ is the correct model is estimated at .92, which is high, but not high enough to make a reliable choice of dependency direction.

Rows 4–8 of Table 1 present a more fine-grained analysis of the influence of individual dependency parameters on G^2. The results

suggest that c_0 (boundary position dependency) is the most important dependency parameter, followed by c_α (boundary orientation dependency) and c_β (boundary steepness dependency). A combination of c_0 and either c_α or c_β (rows 7 and 8) produces similar G^2 values.

The bottom row presents the results for refitting Nearey's LR model with secondary cues and diphone bias (Nearey, 1990, Model 4, Table 4). As small differences in practical aspects of model fitting, such as choice of minimization algorithms, may produce differences in G^2 values, a refit of Nearey's model was carried out to ensure a fair comparison with HICAT. Clearly, HICAT Model 4, including only c_0, produces a comparable GOF to Nearey's diphone-biased secondary cue model.

As discussed earlier, the FLMP (Massaro, 1998) is fundamentally unsuitable to model data like those of the *bad bet* experiment because it makes the implicit assumption that there is no cue sharing among the two categorizations. This assumption obviously does not hold in the *bad bet* case, nor in the other experiments discussed below. Therefore, FLMP model fits were not included in the present analyses.

Table 2 lists statistical tests of the significance of differences in GOF between several models (refer to Nearey, 1990, for a full description of the mechanics of the quasi-likelihood tests). The top row of Table 2 shows a comparison of the independent model and model $V \rightarrow C$ when it includes only the position dependency parameter c_0. The improvement is highly significant. Adding either boundary orientation parameter c_α or boundary steepness parameter c_β to this model produces small but significant improvements in the fit. Adding the yet remaining parameter (bottom row) does not further increase GOF, however. Although Model 8 produces a slightly lower G^2 than Model 7, it is impossible to make a well-founded choice between the two on the basis of further significance tests because they are not nested and have equal numbers of parameters. Model 8, that is, the model in which the position and steepness of the /d/–/t/ boundary depends on the preceding vowel, is therefore tentatively selected as the best-fitting model, with Model 7 a close second.

Figure 2 presents the territorial plot associated with Model 8. Unfortunately, the steepness dependencies are only visually explicit in the actual probability surfaces, not in territorial plots like Figure 2. A comparison of Figure 2 to Nearey's (1990) Figure 3 reveals that the two territorial plots are very similar. *Bat* and *bed* occupy most of the territory, and they share a small boundary segment. *Bad* and *bet*, on the other hand, do not meet. There are a

Table 2
Quasi-Likelihood Tests for Various Models for the Categorization Data of the Bad Bet Experiment by Whalen (1989)

Model comparison	Extra parameters	ΔG^2	$\frac{\Delta G^2}{G^2_{max}}$ (%)	Δdf	F ratio	p
4–1	c_0	65	29	1	33.5	<.001
7–4	c_α	16	10	1	9.0	<.005
8–4	c_β	17	11	1	9.7	<.005
2–8	c_α	5	3.5	1	2.9	*ns*

Note. For interpretation of model numbers, refer to Table 1.

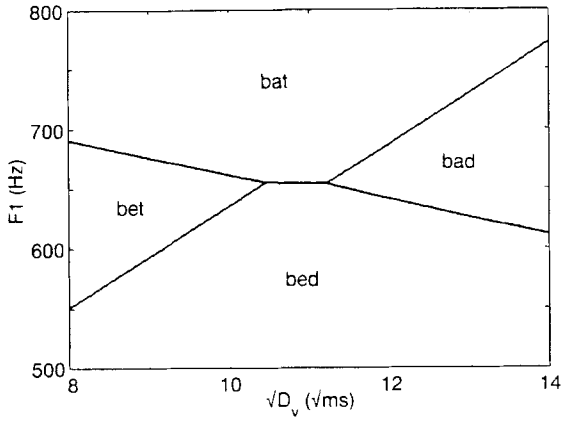

Figure 2. Territorial plot associated with the best-fitting HICAT model (Model 8, Table 1) for Whalen's (1989) *bad bet* experiment. In this model, the position and steepness of the /d/–/t/ boundary depends on the perceived vowel. $\sqrt{D_v}$ represents the square root of the vocoid duration. Compare this figure to Figure 3 in Nearey (1990).

number of differences between the two plots, however. First, in Nearey's plot, indeed in all territorial plots within the LR class (without cue interactions), category boundaries are straight lines. This is not the case for HICAT. The syllable boundaries associated with the primary categorization (in this case the *bed–bad* boundary and the *bet–bat* boundary) are nonlinear when parameters c_0 and/or c_β are nonzero. The cause of this nonlinearity is explained in Smits (in press). The syllable boundaries associated with the secondary categorization (in this case the *bed–bet* and *bad–bat* boundaries), on the other hand, are linear and parallel, like in Nearey's fit.

Second, beside the boundary position dependency, the HICAT model illustrated in Figure 2 contains a steepness dependency: the *bad–bat* boundary is steeper than the *bed–bet* boundary. This steepness dependency means that if one makes a cross-section of psychological space perpendicular to the *bad–bat* boundary, the /d/–/t/ categorization functions are steeper (i.e., change more rapidly) than along a similar cross-section perpendicular to the *bed–bet* boundary. Why did Nearey's fit not reveal such dependency? First of all, the LR model simply does not incorporate hierarchical dependencies, although it could in principle model some form of (nonhierarchical) steepness dependency. However, because of the differences in the mathematical structure of the two models, modeling a steepness dependency in LR would require several free parameters, whereas in HICAT it requires only one.

Sushi Experiment

For Experiment 3, Whalen (1989) created a 10-member synthetic fricative–vowel continuum by orthogonally varying the frequency of a fricative resonance F_{fr} in three steps and $F2$ in four steps, leaving out two of the resulting stimuli. Listeners were asked to classify the stimuli as either *see, Sue, she,* or *shoe.* Each stimulus was presented 20 times to each of 10 listeners, and results were pooled across listeners.

Again, the HICAT model was initially fitted three times: once as the independent model ($c_0 = c_\alpha = c_\beta = 0$), once assuming the

vowel categorization depends on the fricative ($F \rightarrow V$), and once assuming the opposite hierarchy ($F \leftarrow V$). The modeling results are given in rows 1, 3, and 4 of Table 3. Making the vowel categorization independent of F_{fr} produced very little change in GOF for model $F - V$; see row 2 of Table 3. As in the *bad bet* experiment, the introduction of categorization dependencies greatly improves the GOF compared with the independent model (see rows 1, 3, and 4). Model $F \leftarrow V$ has a slight advantage over model $F \rightarrow V$, suggesting that the fricative categorization depends on the vowel, and not vice versa. The choice of direction is, however, very unreliable. The probability of choosing the correct direction is estimated at .55, close to chance level. This low reliability is mainly caused by the lack of data.

The tests for significance of model parameters gave very straightforward results, and the equivalent of Table 2 for the *sushi* experiment is therefore left out to save space. Of the three dependency parameters, the boundary position parameter c_0 produced by far the largest drop in G^2. As was the case for the $F - V$ model, making the vowel categorization independent of F_{fr} produced no significant increase in G^2 (Model 5). Adding either c_α or c_β to Model 5 in Table 2 produced no further significant decrease of G^2, so Model 5 was selected as the best model. Row 6 shows the GOF of Nearey's (1990) diphone-biased secondary cue model, excluding the interaction between vowel response and F_{fr} (Nearey, 1990, Table 8, Model 4). Again, the G^2 values for our selected Model 5 and Nearey's Model 4 are very similar. Both models have six free parameters.

Figure 3 shows the territorial plot of Model 5. The main difference with Nearey's corresponding plot (Nearey, 1990, Figure 5) is that in his plot the /su/–/si/ and /ʃu/–/ʃi/ boundaries are parallel straight lines, whereas in the HICAT plot these boundaries are nonlinear and asymptotically tend to the same vertical line.

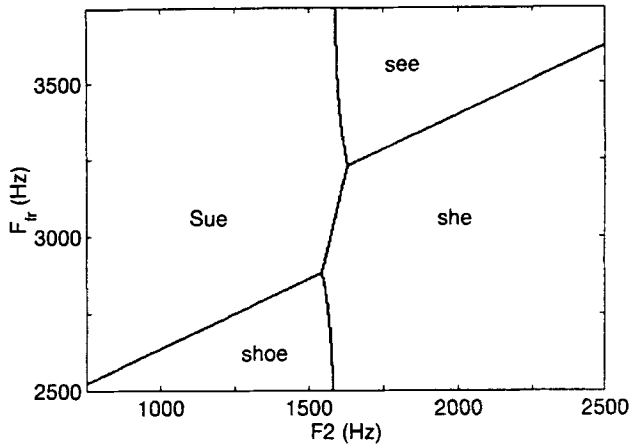

Figure 3. Territorial plot associated with the best-fitting HICAT model (Model 5, Table 3) for Whalen's (1989) *sushi* experiment. In this model, the position of the /s/–/ʃ/ boundary depends on the perceived vowel. Compare this figure to Figure 5 in Nearey (1990).

Discussion

The *bad bet* and *sushi* experiments by Whalen (1989) address a fundamental problem in speech perception research: How do listeners decode coarticulated speech? The *bad bet* case is interesting because it is an extreme example of cue sharing: Both vowel height and final stop voicing are known to be strongly affected by vocoid duration. The *sushi* experiment is interesting because it constitutes a case where listeners have to deal with strong coarticulatory effects on a relatively stationary portion of the acoustic signal. Because of spreading of the feature (round) from the vowel to the preceding fricative, the fricative spectrum is highly affected by the vowel. Whalen's experiments showed that listeners' categorizations of successive phonemes cannot be independent in the strongest sense; that is, the categorization process must involve cue-sharing, phonological context effects, or both. Nearey's (1990) LR analyses and the present HICAT analyses provide similar but not identical accounts of Whalen's data. Both Nearey's analyses and the HICAT analyses support the hypothesis that listeners share cues between phoneme categorizations. The cue-sharing hypothesis is not enough for a full account of Whalen's data, however. In Nearey's model a diphone bias is used to model the finding that in regions of psychological space where four categories meet, responses associated with two diagonally opposed regions may be more likely than the other two. The HICAT model suggests an alternative mechanism in the form of a categorization hierarchy: One categorization depends on the other, but not vice versa. A similar effect to that of Nearey's diphone bias is created by a boundary position dependency in HICAT. Besides the boundary position dependency, HICAT offers two more possible types of dependency associated with steepness and orientation of the category boundary. In the analysis of Whalen's *bad bet* data, the steepness dependency proved to make a significant contribution.

Why do listeners use strategies involving these dependencies? The theory presented by Smits (in press) proposes an explanation in terms of distributions of acoustic cues in natural speech. Listeners have tuned their phoneme classifiers to these cue distributions. If the distributions of cues signaling a particular distinction are more overlapping in Context 1 than in Context 2 (a case of the converged geometry), the category boundary for the distinction should optimally be more shallow, or uncertain, in Context 1 than in Context 2. The theory would therefore predict for the current case that the two-dimensional distributions of D_v and F1 in naturally produced tokens of the words *bad* versus *bat* are better

Table 3
Goodness-of-Fit Analysis for Various Models for the Categorization Data of the Sushi Experiment by Whalen (1989)

Model	Parameters	G^2	RMSD (%)	*df* res.	Overdisp.
1	$F - V$, full	270	7.32	24	11.25
2	$F - V - V(F_{fr})$	274	7.57	25	10.96
3	$F \rightarrow V$, full	30.6	2.56	21	1.46
4	$F \leftarrow V$, full	29.6	2.31	21	1.41
5	$F \leftarrow V$, $c_0 - V(F_{fr})$	36.5	2.71	24	1.52
6	LR $- V(F_{fr})$ + diph.	36.2	2.55	24	1.51

Note. $F - V$ refers to the independent model, containing six parameters p_0, p_x, p_y, q_0, q_x, q_y; $F \rightarrow V$ refers to the model in which the vowel categorization depends on the fricative; $F \leftarrow V$ refers to the model with the opposite hierarchy. $- V(F_{fr})$ indicates that the vowel categorization is not depending on F_{fr}; c_0 indicates that c_0 was the only dependency parameter included; LR + diph. refers to Nearey's (1990) logistic regression model, which includes a diphone bias but no stimulus-tuned diphone terms. RMSD = root mean square difference; *df* res. = residual degrees of freedom; Overdisp. = overdispersion.

separated than those for *bed* versus *bet*. Similarly, the fact that there is a significant boundary position dependency (c_0 is nonzero) suggests that the cue distributions show a shifted pattern. Verification of these predictions is a matter for future research.

For the *sushi* data, only a position dependency was found, which is associated with a shifted geometry of the associated cue distributions. This means that F_{fr} values for /s/ and /ʃ/ are generally higher in frequency when followed by /i/ than when followed by /u/. Although there are no quantitative data sets across many speakers available to confirm this, the qualitative pattern is well known and indeed was used by Whalen (1989) as a motivation to study the recognition of fricative–vowel syllables. However, the earlier acoustic–phonetic analysis of the influence of lip rounding on the fricative spectrum suggested that the rounding-induced frequency shift of the main resonance would be larger for /s/ than for /ʃ/. If this analysis is accurate, an additional convergence pattern would occur, which would predict a potential steepness dependency if the fricative categorization depends on the vowel or an orientation dependency for the reverse hierarchy (see Smits, in press). The HICAT analyses presented above do not show such dependencies. However, as mentioned earlier, the data set is very small. Presence of such steepness dependencies is tested in the larger experiment on categorization of Dutch syllables /si sy ʃi ʃy/ presented in a later section.

The *sushi* experiment and its analyses remain somewhat unsatisfactory. First of all, the experimental data set is very small, which may have prevented interesting processing mechanisms to reach significance. Second, an explicit link of the processing strategies to statistical distributions of acoustic cues—which forms an essential part of a pattern-recognition approach—cannot be made because the acoustic data are not available. Therefore, the present study aimed to replicate Whalen's (1989) *sushi* experiment on a larger scale and also to complement the perception data with a set of acoustic measurements.

The Dutch phoneme inventory allows a phonologically somewhat cleaner and more symmetrical version of the *sushi* experiment. Besides vowels /i/ and /u/, Dutch has the vowel /y/, with /i/ and /y/ differing only in the feature [round]. Furthermore, in contrast to English, Dutch /ʃ/ is unrounded, so Dutch /s/ and /ʃ/ differ only in place of articulation. As in English, vowel rounding spreads to preceding fricatives (Booij, 1995; Collins & Mees, 1981). On the basis of these considerations, we focused on the production and perception of Dutch syllables /si sy ʃi ʃy/.

Two experiments were conducted. In Experiment 1, acoustic cues relevant to the perception of the fricative and vowel distinctions of interest were measured on a large set of natural utterances. On the basis of these utterances, the most likely dependency direction and type were predicted, assuming that listeners operate as statistical pattern recognizers. Experiment 2 tested these predictions by presenting a two-dimensional synthetic fricative–vowel continuum to listeners for categorization and analyzing the data for processing dependencies using the HICAT model. The simplifying assumption was made that as long as there is no variation in speaker sex, a listener's training material is well approximated by a set of cue distributions of many speakers mixed together. Accordingly, the talkers in Experiment 1 were all adult men, and the synthetic stimuli in Experiment 2 emulated a male voice.

Experiment 1

Method

Participants. Seventeen adult male talkers participated in the experiment, none of whom had a history of speech or hearing problems.

Procedure. Participants were seated in a soundproof booth in front of a microphone and a computer screen. For each participant a different random list was constructed of three repetitions of each of the syllables /si sy ʃi ʃy/. In accord with Dutch spelling, the orthographic representation of the syllables was *sie, suu, sjie, sjuu*. The syllables were successively presented on the computer screen, and the participants were asked to pronounce them at a comfortable level and rate. Participants were tested individually.

The utterances were recorded on DAT tape and were subsequently converted to digital speech files sampled at a rate of 16 kHz using a low-pass filter with a cutoff frequency of 7.5 kHz.

Acoustic measurements. A blind measurement procedure was adopted; that is, the experimenter did not know what syllable or speaker he was doing measurements on. To this end, a randomized list of the 204 syllables was created, and all speech files received a name that only contained its position on the list.

The acoustic measurements were designed to capture the dominant acoustic dimensions associated with the phonetic contrasts under study. Figure 4 gives spectra from relatively stationary portions of the frication noise and the vowel of characteristic tokens of /si sy ʃi ʃy/ uttered by one of the participants. Dutch /s/ is different from English /s/ in being more laminal and less tense. Furthermore, Dutch /ʃ/ differs from English /ʃ/ in being unrounded (Collins & Mees, 1981). As can be seen in the left-hand column of Figure 4, fricative spectra of all four syllables have the following characteristics. At mid to high frequencies (between 2 and 5 kHz), there is a spectral peak (indicated by the dashed line in Figure 4). To the right of this peak, there is a more or less flat plateau of high energy, whereas to the left of the peak, there is relatively little energy. The frequency of the spectral peak was chosen as the dominant parameter associated with the /s/–/ʃ/ contrast. This parameter, henceforth indicated as F_{fr}, was measured by hand on wideband spectrograms. Care was taken that the measurements were made in relatively stationary portions of the signals, not in transitional regions. Figure 4 clearly illustrates how F_{fr} varies with fricative as well as vowel identity.

According to Stevens (1998), lip rounding for nonlow front vowels such as /i/ has the following acoustic consequences: lowering $F2$ and $F3$, bringing $F2$ and $F3$ closer together (in other words, $F3$ is lowered more than $F2$), and decreasing the bandwidth of the third formant. Pols, Tromp, and Plomp (1973) measured formant frequencies and amplitudes on Dutch vowels in /hVt/ context, spoken by 50 male talkers. Their analysis showed that in accordance with Stevens's claims, the Dutch /i/–/y/ contrast correlates with frequencies of $F2$ and $F3$, both being lower for /y/. Pols et al. (1973) reported mean $F2$ and $F3$ frequencies of 2208 Hz and 2766 Hz for /i/, and 1730 Hz and 2208 Hz for /y/, respectively. Thus, the downward frequency shift due to rounding was smaller for $F2$ (478 Hz) than for $F3$ (558 Hz), but the difference was not large.

For the purpose of our experiment, both $F2$ and $F3$ were measured on wideband spectrograms of all utterances. Like F_{fr}, $F2$ and $F3$ were measured in relatively stationary parts of the speech signal. Mean frequencies for $F2$ and $F3$ were 2088 Hz and 2795 Hz for /i/, and 1762 Hz and 2175 Hz for /y/, respectively. These values compare well with those found by Pols et al. (1973), except for $F2$ for /i/, which is 120 Hz lower in the present data. The right-hand panels of Figure 4 illustrate the frequency shifts in $F2$ and $F3$ due to rounding and show furthermore that the vowel spectra are relatively insensitive to the identity of the preceding fricative. For the present data, the frequency shift due to rounding is much larger for $F3$ (620 Hz) than for $F2$ (326 Hz), which would favor $F3$ over $F2$ as a cue for the rounding contrast. In addition, linear discriminant analysis for /i/ and /y/ using dimensions $F2$ and $F3$ showed that an optimal linear classifier would

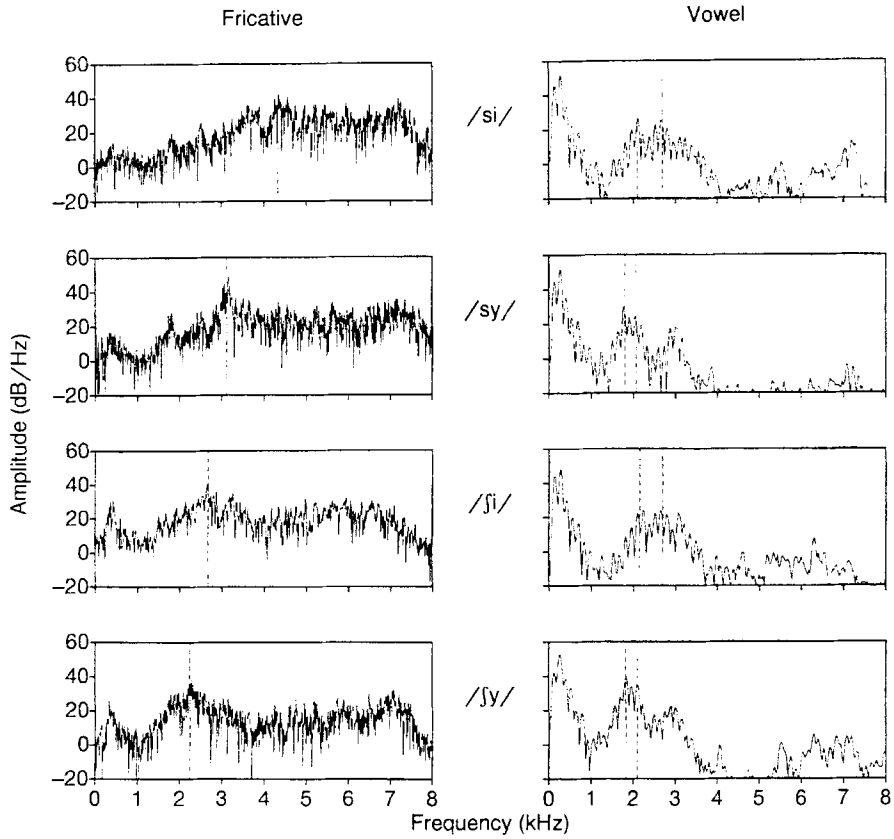

Figure 4. Characteristic spectra taken from relatively stationary portions of the fricative (left-hand column) and vowel (right-hand column) of syllables /si sy ʃi ʃy/, spoken by one of the participants of Experiment 1. The dashed lines in the fricative spectra indicate the frequency F_{fr} of the fricative resonance. The dashed lines in the vowel spectra indicate the frequencies of the second and third formants, respectively. Note how F_{fr} varies with both fricative and vowel, whereas the vowel formants are more or less insensitive to fricative identity.

assign a weight to *F*3 that is four times larger than the weight for *F*2 in the /i/–/y/ classification. On the basis of these results, *F*3 was selected as the dominant acoustic dimension associated with the /i/–/y/ contrast.

Results and Discussion

Figure 5 represents a scatterplot of the measurements of F_{fr} and *F*3 on the set of 204 utterances, plotted on linear frequency scales (A) and on ERB scales (B), which is a tonotopic frequency scale (Glasberg & Moore, 1990). Panel B is assumed to correspond to the psychological space of listeners in the perception experiment below. Table 4 reports the means and (co)variances of the two-dimensional syllable PDFs fitted to the measurements. Covariance matrices were calculated separately for different syllables.

Figure 5 and Table 4 reveal several interesting aspects of the acoustic data that listeners have to deal with when recognizing fricative–vowel syllables. The first striking aspect is that although 17 different speakers produced the utterances, there is relatively little overlap between the distributions of the four syllables. Of course, the syllables were produced in isolation in laboratory circumstances, which would be expected to keep variability to a minimum. Corresponding measurements from conversational speech would probably display more overlap between syllables.

Second, the covariance matrices for the four syllables are not very different. Compare, for example, the variability among ellipses in Figure 5 to the variability among ellipses for vowel formants of different American English vowels, as estimated by Hillenbrand, Getty, Clark, and Wheeler (1995). The largest difference in the standard deviations is the standard deviation of F_{fr} being roughly 1.5 times as big for /si/ as for /su/. There are more substantial differences in the correlation coefficients, but it is striking that all coefficients are positive. (The same effect occurred in Hillenbrand et al., 1995.) This effect is probably related to vocal tract length variability. Speakers with large vocal tracts will generally produce lower resonance frequencies for both fricative and vowel spectra than speakers with short vocal tracts. Taken together, the assumption of equal covariance matrices for the four syllables that was made for the predictions below is consistent with the observations.

With respect to the means of the syllable PDFs, Figure 5 and Table 4 reveal that as expected, the influence of the vowel on F_{fr} is much larger than the influence of the fricative on *F*3. The effects of vowel rounding on F_{fr} are both a downward shift in frequency and a convergence; that is, F_{fr} means are lower for fricatives in context /y/ than /i/, and the distance between F_{fr} means for /s/ and

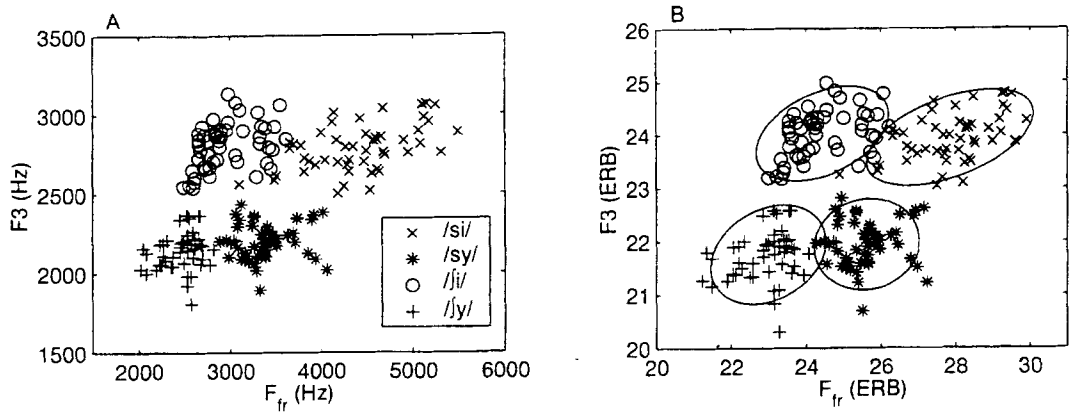

Figure 5. Scatterplots of the values of acoustic cues F_{fr} and *F*3 measured on 51 tokens of each of the syllables /si sy ʃi ʃy/ in Experiment 1. A: Data on linear frequency scales; B: Data on perceptually more plausible equivalent rectangular bandwidth (ERB) scales. In addition, ellipses that are plotted in B represent isoprobability contours of two-dimensional Gaussian probability-density functions, estimated separately for each syllable.

/ʃ/ is smaller for /y/ than for /i/ (see Table 4). Thus, a pattern much like that of Figure 1D is created. In contrast, the *F*3 means for /si/ and /ʃi/ are very similar, as are those for /sy/ and /ʃy/ (see Table 4). Consequently, a vowel categorization based only on *F*3 and independent of the fricative is close to optimal (horizontal category boundary in Figure 5). For a good fricative categorization, on the other hand, both F_{fr} and *F*3 are needed (slanted boundary in Figure 5), and a dependency of the fricative categorization on the vowel may be called for.

The prediction method defined in Smits (in press) was applied to the acoustic measurement data presented above to predict the dependency direction and type most likely to be used by listeners in the recognition of syllables /si sy ʃi ʃy/. Recall that the prediction method essentially translates the four dependency types (independence, boundary position, orientation, and steepness dependence) into four sets of conditions on the phoneme PDFs. Using these conditions, inferred syllable PDFs as well as the optimal boundaries between them can be calculated. Figures 6A–6D show isoprobability contours (ellipses) of the inferred syllable PDFs for assumptions of independence (A), position dependence (B), orientation dependence (C), and steepness dependence (D). For B, C, and D, it was assumed that the fricative categorization depends on the vowel. Figure 6E shows the actual syllable PDFs, assuming equal covariance matrices. In addition to the isoprobability contours, the syllable boundaries are shown, as well as the actual acoustic data.

Table 4
Means, Standard Deviations, and Correlation Coefficients of F_{fr} and F3 for the Syllables /si sy ʃi ʃy/

Syllable	$M(F_{fr})$	$M(F3)$	$SD(F_{fr})$	$SD(F3)$	R
/si/					
Hz	4386	2781	545	146	0.51
ERB	27.85	23.94	1.12	0.45	0.50
/sy/					
Hz	3384	2199	280	111	0.08
ERB	25.62	21.94	0.70	0.43	0.07
/ʃi/					
Hz	2958	2808	309	145	0.38
ERB	24.44	24.02	0.88	0.45	0.41
/ʃy/					
Hz	2495	2151	223	118	0.27
ERB	22.99	21.75	0.77	0.46	0.27

Note. ERB = equivalent rectangular bandwidth.

The prediction method produces an ordering of dependency types according to decreasing likelihood of occurrence on the basis of a comparison of inferred syllable PDFs (Figures 6A–6D) with the actual ones (Figure 6E). The dissimilarities between the inferred and actual PDFs are expressed in terms of the Bhattacharyya distance (Fukunaga, 1990). If the fricative categorization depends on the vowel, then the following ordering of dependency types is found (and associated average Bhattacharyya distances) in decreasing order of likelihood of occurrence: steepness dependence (.0047), position dependence (.0135), orientation dependence (.0137), and independence (.0158). It can be visually verified from Figure 6 that among Figures 6A–6D, the ellipses of 6D (steepness dependency) are most similar to those of 6E. Introducing a steepness dependency leads to a substantial reduction in PDF dissimilarity (from .0158 to .0047) and therefore also a substantial improvement in (fuzzy) classification performance. Position or orientation dependencies, on the other hand, produce only small improvements that do not differ much in size.

So far, it was assumed that the fricative categorization depends on the vowel. For the reverse hierarchy, that is, the vowel categorization depending on the fricative, the following ordering is found: orientation dependence (.0058), position dependence (.0135), steepness dependence (.0144), and independence (.0161). If we compare the predictions for the two dependency directions, two things become obvious. First, the hierarchy involving a dependency of the fricative categorization on the vowel seems to be more useful than the reverse dependency, based on the Bhattacharyya distances for the best-performing dependency types. The difference, is, however, small (.0047 vs. .0058). Second, in terms of the underlying geometry of cue distributions, the orderings of dependency types for the two dependency directions are in perfect

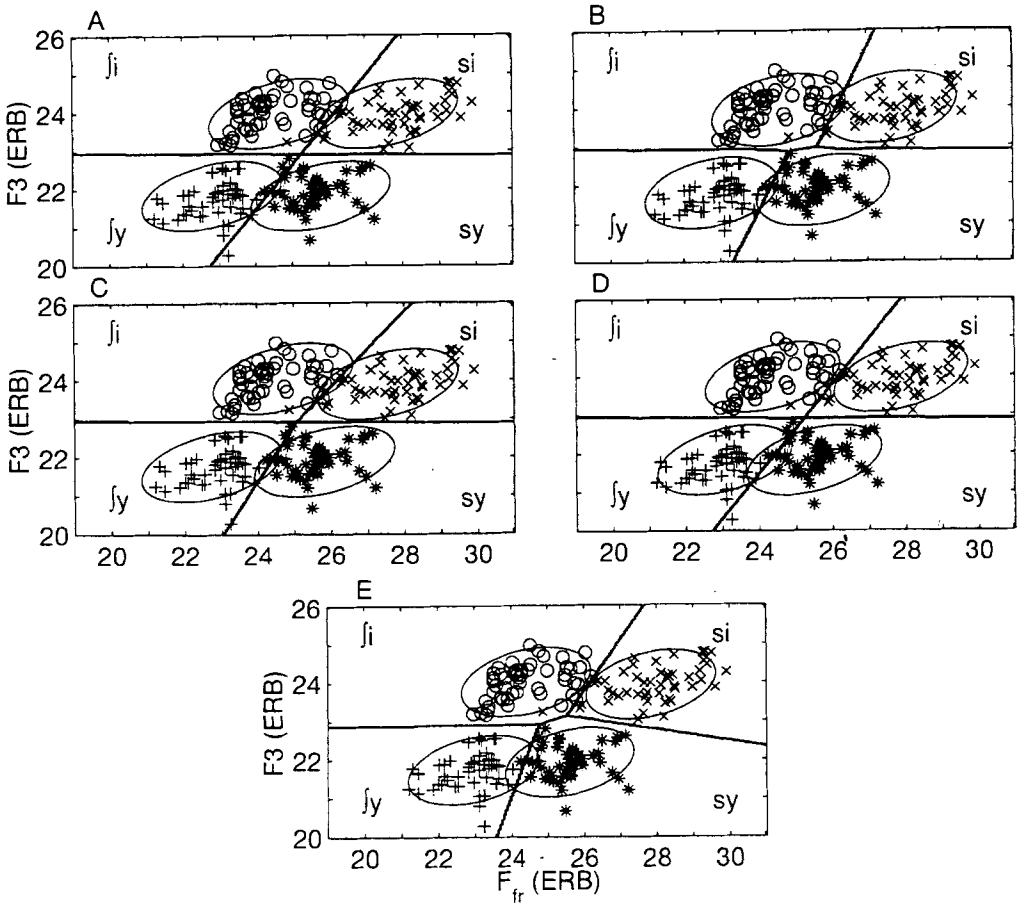

Figure 6. Inferred syllable probability-density functions (PDFs; ellipses) and category boundaries (solid lines) predicted from the acoustic data of Experiment 1, assuming independence (A), boundary position dependence (B), orientation dependence (C), and steepness dependence (D). For B–D, it was assumed that the fricative categorization depends on the vowel. E: Syllable distributions and boundaries assuming syllable-based recognition and equal covariance matrices. Among A–D, the syllable PDFs of D are most similar to those of E in terms of the Bhattacharyya distance. ERB = equivalent rectangular bandwidth.

agreement. Recall that a shifted PDF geometry is associated with position dependency for both dependency directions, whereas the convergence geometry is associated with steepness dependency for one direction and orientation dependency for the other.

In summary, then, it is predicted that (a) listeners are more likely to use a categorization hierarchy in which the fricative categorization depends on the vowel and (b) if listeners use such strategy, it is the steepness parameter of the fricative boundary that will show the strongest dependency. If, on the other hand, listeners use the reverse dependency direction, orientation dependency is the most likely.

Experiment 2

In perception experiments such as the present one, the experimenter is confronted with a choice between focusing on group data, that is, data pooled across many listeners, or on data from individual listeners. Although the first option is generally more popular, for the present research the second option was chosen. The reason for this choice is that at present, it is unclear what the effect of summing data across listeners is on the categorization strategies that are inferred from subsequent HICAT analyses. It is conceivable that data pooling would erase or even introduce apparent categorization dependencies in the group data that are (not) present in the individual data. Therefore, I decided to use a relatively small group of 4 listeners and collect a large set of categorizations per listener. Group analyses remain, of course, possible for such a data set and are compared with the individual sets below.

Method

Participants. Two male and two female students at Nijmegen University participated in the experiment. None had any history of hearing or speech difficulties.

Stimuli. A set of 64 stimuli was synthesized using the Praat software (Boersma & Weenink, 1999) by orthogonally varying the frequency of a fricative pole F_{fr} and *F*3 of a vocalic portion in eight steps. Note that these synthesis parameters correspond to the acoustic dimensions measured on the natural utterances in Experiment 1. As in Whalen's (1989) Experiment 3, stimuli consisted of a fricative portion and a vowel portion spliced

together. Also as in Whalen's experiment, it was decided to zoom in on the most interesting region of psychological space, that is, a region where the stimuli are relatively ambiguous. The excitation signal for the fricative portion was a preemphasized Gaussian white noise of 180 ms. The excitation signal was filtered by a cascade formant filter using two formants. One formant was fixed at 6500 Hz (2600 Hz bandwidth), and the other had a frequency that varied between 2890 and 3310 Hz in eight steps of 60 Hz. The bandwidth of this formant was 10% of its frequency. The amplitude envelope of the frication noise linearly rose from zero to its maximum over the first 100 ms of the signal, remained at its maximum for 20 ms, and linearly fell to zero over the last 60 ms.

The excitation signal of the vocalic portion of the syllable consisted of a 250-ms pulse train with *F*0 linearly falling from 160 to 120 Hz. To increase naturalness of the resulting vowel sounds, amplitudes of the first and second harmonics of the excitation signal were multiplied by a factor of 6 or 3, respectively. The envelope of the excitation signal linearly rose from one fifth of its maximum value to its maximum value over the first 31 ms, remained at its maximum during the next 119 ms, and then linearly fell to zero over the final 100 ms. The excitation signal was filtered by a cascade formant filter consisting of seven formants. A convincing Dutch /i/–/y/ continuum could be created by varying only *F*3, setting *F*2 at a compromise value. Within stimuli, frequencies and bandwidths of all formants were stationary. *F*3 varied between stimuli from 2450 to 2625 Hz in eight 25 Hz steps, with a constant bandwidth of 110 Hz. Frequencies (*F*) and bandwidths (*B*) of other formants were the same for all stimuli: *F*1 = 250 Hz, *B*1 = 40 Hz; *F*2 = 1800 Hz, *B*2 = 80 Hz; *F*4 = 3200 Hz, *B*4 = 130 Hz; *F*5 = 4000 Hz; *B*5 = 150 Hz; *F*6 = 5500 Hz, *B*6 = 200 Hz; *F*7 = 6500 Hz, *B*7 = 200 Hz. These values were inspired by Scheffers (1983).

The resulting eight fricative and eight vocalic segments were orthogonally combined to form 64 fricative–vowel syllables.

Procedure. The experiment consisted of four sessions. Session one was a familiarization session in which each of the 64 stimuli was presented five times to each listener in randomized order. Listeners were asked to classify stimuli as *sie, suu, sjie,* or *sjuu* (Dutch orthographic representations of /si sy ʃi ʃy/, respectively) by pressing one of four buttons on a response box. The ordering of the button labels was changed between listeners. The data of the familiarization session were discarded.

Next, listeners were presented with 60 repetitions of each of the 64 stimuli. Because of the total duration of the experiment, stimuli were presented in three sessions of 1,280 trials each. Stimuli were completely randomized within a session, and different randomizations were used in different sessions and for different listeners. The task of the listeners was the same as in the familiarization session.

Stimulus presentation was controlled by a PC. Listeners were tested individually in a sound-treated booth. Stimuli were presented over headphones at a comfortable level. In all sessions, listeners had 2,070 ms to respond after stimulus offset. If no response was given in this period, no response was collected for the trial, and the next stimulus was presented. Such failures to respond occurred only on a very small proportion of trials (0.25%). Listeners got a short break after every 100 trials.

Results

Group analyses. For the first analyses of the categorization data, results were pooled across participants, yielding 240 responses for each of the 64 stimuli.

Before doing the HICAT analyses, I checked whether the stimulus continuum had been successful in cueing the fricative and vowel contrasts through the variation of the stimulus parameters F_{fr} and *F*3. To this end, I calculated average fricative categorizations for stimuli with F_{fr} at its extreme values. In analogous fashion, I calculated average vowel categorizations for stimuli with *F*3 at its extreme values. The results showed that stimuli with minimal F_{fr} (2890 Hz) received a label containing the fricative /ʃ/ in 83% of the cases, whereas stimuli with maximal F_{fr} (3310 Hz) were labeled /s/ 82% of the time. For the vowels, it was found that stimuli for which *F*3 had its lowest value (2450 Hz) were given a label containing the vowel /y/ 96% of the time. Stimuli with maximum *F*3 (2625 Hz) were labeled /i/ 92% of the time. Given that the stimulus continuum was relatively small compared with the means of *F*3 and F_{fr} found for the natural utterances, it can be concluded that the selected acoustic parameters successfully cued the respective phonetic distinctions and that convincing endpoints were created.

Next, the HICAT model analyses were carried out. Like in the reanalyses of Whalen's (1989) data, the HICAT analyses started with three model fits: the independent model ($F - V$), the full dependency model with the vowel categorization depending on the fricative ($F \rightarrow V$), and the full dependency model with the fricative categorization depending on the vowel ($F \leftarrow V$). For the model fits, the two acoustic parameters varied in the stimuli (F_{fr} and *F*3) were recoded in terms of ERB. Results of the HICAT analyses are presented in Table 5.

Both dependency models give substantial improvements in G^2 over the independent model. The G^2 difference between models $F \rightarrow V$ and $F \leftarrow V$ is 43, or 11%. The estimation technique given in Smits (in press) estimates the probability of incorrectly choosing the dependency direction of $F \leftarrow V$ at 0.015. Thus, the group data support a hierarchy involving a dependency of the fricative categorization on the perceived following vowel.

Rows 4–9 of Table 5 illustrate the influence of individual model parameters of model $F \leftarrow V$ on G^2. A comparison of Models 4, 5, and 6 shows that of the three dependency parameters, the steepness dependency parameter c_β gives the largest reduction of G^2 compared with the independent model. Row 10 of Table 5 gives the GOF of Nearey's (1990) diphone-biased segment LR model. Its GOF is comparable to model $F \leftarrow V$ including only the position dependency parameter c_0 (Model 4). The similarity in behavior of the diphone-biased LR model and HICAT including only a position dependency was discussed in Smits (in press).

Table 6 gives the outcomes of statistical tests of the significance of the contributions of various model parameters. The top row

Table 5
Goodness-of-Fit Analysis for Various Models for the Pooled Categorization Data of Experiment 2

Model	Parameters	G^2	RMSD (%)	*df* res.	Overdisp.
1	$F - V$, full	432	3.75	186	2.32
2	$F \rightarrow V$, full	389	3.61	183	2.13
3	$F \leftarrow V$, full	346	3.37	183	1.89
4	$F \leftarrow V$, c_0	414	3.69	185	2.24
5	$F \leftarrow V$, c_α	421	3.73	185	2.27
6	$F \leftarrow V$, c_β	385	3.50	185	2.08
7	$F \leftarrow V$, $c_0 + c_\beta$	364	3.42	184	1.98
8	$F \leftarrow V$, $c_\alpha + c_\beta$	362	3.45	184	1.97
9	$F \leftarrow V$, full $- V(F_{fr})$	358	3.45	184	1.95
10	LR + diph. bias	413	3.69	185	2.23
11	LR + diph. bias $- V(F_{fr})$	414	3.69	186	2.23

Note. $F - V$ refers to the independent model; $F \rightarrow V$ refers to the model in which the vowel categorization depends on the fricative; $F \leftarrow V$ refers to the model with the opposite hierarchy. RMSD = root mean square difference; *df* res. = residual degrees of freedom; Overdisp. = overdispersion.

Table 6
Quasi-Likelihood Tests for Various Models for the Categorization Data of Experiment 2

Model comparison	Extra parameters	ΔG^2	$\frac{\Delta G^2}{G^2_{max}}$ (%)	Δdf	F ratio	p
6–1	c_β	47.0	10.9	1	22.6	<.001
7–6	c_0	21.2	5.5	1	10.7	<.005
8–6	c_α	23.6	6.1	1	12.0	<.001
3–8	c_0	15.4	4.3	1	8.1	<.01
9–8	$V(F_{fr})$	12.1	3.4	1	6.4	<.025

Note. For interpretation of model numbers, refer to Table 5.

shows that Model 6 ($F \leftarrow V$ including only steepness dependency parameter c_β) gives a significantly better fit than independent Model 1. Rows 2, 3, and 4 demonstrate that addition of both remaining parameters c_0 or c_α significantly improves the fit still further. Finally, setting the parameter that codes the dependency of the vowel categorization on F_{fr} to zero does not significantly worsen the fit (I decided to set the significance criterion at $p = .01$ rather than $p = .05$ to keep the models more economical). Eventually, Model 9 of Table 5 was selected as the best-fitting model for the pooled categorization data of Experiment 2. This model contains eight parameters. The values of the dependency parameters in Model 9 are $c_0 = -0.101$, $c_\alpha = 0.053$, $c_\beta = 0.133$. It is worth noting that Model 9 shows a significant lack of fit ($p < .001$). This is, however, the rule rather than the exception in the modeling of repeated measures categorization data, especially when pooled across listeners, and does not necessarily indicate the presence of any systematic components not captured by the model (McCullagh & Nelder, 1989). The fit of Model 9 is actually good, with RMSD = .035 and an overdispersion smaller than 2.

Figure 7 presents the territorial plot associated with Model 9. The boundary position dependence creates the small boundary section separating /ʃi/ from /sy/. The dependence of the orientation of the /s/–/ʃ/ boundary on the vowel is reflected in the /sy/–/ʃy/ boundary and the /si/–/ʃi/ boundary being nonparallel. The steepness dependency is not visually represented in Figure 7, but the /si/–/ʃi/ boundary is somewhat steeper than the /sy/–/ʃy/ boundary. Finally, because the vowel categorization does not depend on F_{fr} in Model 9, the boundaries between /si/ and /sy/ and between /ʃi/ and /ʃy/ asymptotically tend toward a horizontal line.

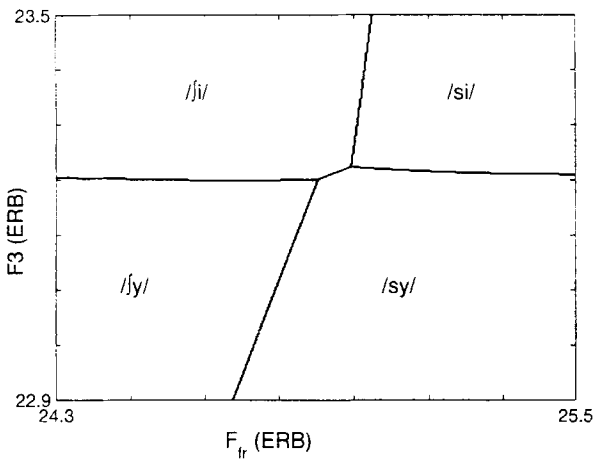

Figure 7. Territorial plot associated with the best-fitting HICAT model (Model 9, Table 5) for the group data of Experiment 2. In this model, position, orientation, and steepness of the /s/–/ʃ/ boundary depend on the perceived vowel. ERB = equivalent rectangular bandwidth.

Individual analyses. Besides the group analysis, the categorization data were analyzed for each listener separately. Table 7 lists for each listener the fits of the full $F - V$, $F \rightarrow V$, and $F \leftarrow V$ models; the fit of Model 9 of Table 5 (the selected model for the group data); and finally, the best-fitting model for the specific subject, along with the estimated reliability of the associated dependency direction. Models were selected on the basis of significance tests, as shown above for the group data, but detailed results of these analyses are not given to save space. All four models provide a good fit to the data, with RMSD values in the order of 5% or less. The fit to the data of Listener 3 is particularly good and is the only one that actually exhibits no significant lack of fit.

The model analyses for 3 out of 4 listeners suggest a categorization hierarchy where the fricative categorization depends on the vowel, as was the case for the group data. Reliabilities of the selection of dependency direction vary between .85 and .99. Oddly, Listener 3's data support the opposite hierarchy, with reliability .94. There seem to be three possible explanations for the lack of consistency of dependency direction among listeners. First

Table 7
HICAT Model Fits for Individual Listeners

Listener and parameters	G^2	RMSD (%)	Overdisp.	Reliability
1				
$F - V$, full	295	5.99	1.58	
$F \rightarrow V$, full	252	5.45	1.38	
$F \leftarrow V$, full	233	5.15	1.28	
$F \leftarrow V$, c_0, c_β, $-V(F_{fr})$	245	5.38	1.33	.94
$F \leftarrow V$, full $-V(F_{fr})$	239	5.31	1.30	
2				
$F - V$, full	370	6.48	1.99	
$F \rightarrow V$, full	343	6.12	1.87	
$F \leftarrow V$, full	301	5.72	1.64	
$F \leftarrow V$, c_α, c_β, $-V(F_{fr})$	304	5.70	1.64	.99
$F \leftarrow V$, full $-V(F_{fr})$	301	5.72	1.63	
3				
$F - V$, full	227	4.82	1.22	
$F \rightarrow V$, full	158	3.68	0.86	
$F \leftarrow V$, full	170	4.34	0.93	
$F \rightarrow V$, c_0, c_α	159	3.74	0.86	.94
$F \leftarrow V$, full $-V(F_{fr})$	175	4.52	0.95	
4				
$F - V$, full	246	5.92	1.32	
$F \rightarrow V$, full	239	5.81	1.31	
$F \leftarrow V$, full	228	5.56	1.25	
$F \leftarrow V$, c_β	232	5.62	1.26	.85
$F \leftarrow V$, full $-V(F_{fr})$	237	5.65	1.29	

Note. $F - V$ refers to the independent model; $F \rightarrow V$ refers to the model in which the vowel categorization depends on the fricative; $F \leftarrow V$ refers to the model with the opposite hierarchy. The fourth row for each listener refers to the best-fitting model, with the final column giving the reliability of the corresponding dependency direction. The fifth row refers to the fit of Group Model 9 of Table 5 to the data of this particular listener. RMSD = root mean square difference; Overdisp. = overdispersion.

of all, the conclusion of the dependency direction may be incorrect for Listener 3. Because categorization of ambiguous stimuli is a stochastic process, it is possible that by chance the data may favor one hierarchy, although the opposite hierarchy was actually used. The probability of this contingency is estimated at .06, which is small but not impossible. A second possibility is that during acquisition of their speech-perception abilities, listeners may have settled on different strategies. This possibility may be due to different input or to some listeners arriving at less effective strategies than others. Nevertheless, one would expect differences to be restricted to relatively detailed and continuous-valued aspects of the categorization process, such as the precise location of categorization boundaries. Finally, it is possible that the fundamental assumption of hierarchical categorization does not apply to the processes under study. For example, mutual rather than one-way dependencies between fricative and vowel categorization may be involved. At present a method for robustly distinguishing between the two alternative strategies is not available.

If we inspect the types of dependencies that are included in the selected models, there is again some variability. For 2 out of 4 listeners, the vowel categorization does not depend on F_{fr}, like in the group analysis. The 3 listeners for whom the fricative categorization depends on the vowel all have a significant contribution of the steepness dependency parameter c_β, whereas the best model for Listener 3 includes the orientation dependency parameter c_α. As was explained in Smits (in press), a converged geometry of cue distributions is associated with a steepness dependency for one dependency direction and with an orientation dependency for the opposite direction. Therefore, all selected models support a categorization strategy associated with the same underlying acoustic pattern, that is, one where the overlap between the distributions of F_{fr} for /s/ and /ʃ/ is different for the vowels /i/ and /y/. In this light, it is not surprising that the group analysis indicated that the steepness dependency parameter c_β is the most important one because it is shared by all listeners.

To shed more light on the relation of the individual model analyses and the group analysis, model parameters were compared across participants. First, the full $F \leftarrow V$ model was fitted to the data of each individual listener. Next, the values of each of the nine model parameters for the 4 participants were inspected. Two-tailed t tests were carried out on the model parameters to test whether the mean value of this parameter across participants was significantly different from zero. Although, given the small number of participants, one must be cautious in drawing conclusions from these tests, the resulting p values do provide some additional insight.

First, the analyses revealed that all coefficients of the stimulus terms ($p_0, p_1, p_2, q_0, q_1, q_2$) are significantly different from zero at the $p = .1$ level, except the coefficient of F_{fr} for the vowel categorization. This tallies well with the group analysis, where F_{fr} was found to make no significant contribution to the vowel categorization. Next, among the dependency parameters c_0, c_α, and c_β, only steepness dependency parameter c_β turned out significant at the $p = .1$ level. In fact, the reliability of average c_β being different from zero was $p = .04$. This result strengthens the conclusion that was drawn from the group analysis that the steepness dependency is the dominant dependency here. At the same time, however, it casts some doubt on the reliability of the position (c_0) and orientation (c_α) dependencies incorporated in the selected group model. Nevertheless, apart from this small discrepancy, the analyses show that on the whole the group analysis provided a good summary of the categorization strategies used by the individual group members.

Discussion

Let us start with the most important point: a comparison of predicted and observed categorization strategies. On the basis of the acoustic data of Experiment 1, it was predicted that (a) a hierarchy in which the fricative categorization depends on the vowel is more likely than the reverse hierarchy; (b) if the fricative categorization depends on the vowel, the most likely dependency type is a steepness dependency, whereas for the reverse hierarchy an orientation dependency is most likely. These predictions were borne out to a large extent. First of all, HICAT analysis of the group data strongly supported a hierarchy in which the fricative categorization depends on the vowel. Individual listener analyses showed that the same was true for 3 out of 4 listeners, whereas 1 listener seemed to have used the reverse hierarchy. Second, in the group analysis, allowing a steepness dependency in the categorization led to a much greater improvement in GOF than allowing either a position or an orientation dependency. This result was further supported by the individual analyses. The best-fitting models for all 3 listeners who showed the predicted dependency direction included the steepness dependency parameter c_β. The best-fitting model for the listener who showed the reverse hierarchy contained the orientation dependency parameter c_α, which is related to the same underlying acoustic pattern.

Besides the directions and types of dependencies between the two categorizations, it is interesting to compare other aspects of the categorization process, such as the overall location, orientation, and steepness of the phoneme boundaries. I do this only qualitatively. Figure 8 shows three-dimensional representations of the response surfaces predicted from Experiment 1 (8A) and measured in Experiment 2 (8B). These figures are associated with the territorial plots of Figures 6D and 7, respectively. On the whole, predicted and observed response surfaces are remarkably similar in terms of steepness of the categorization functions, and the locations and orientations of the category boundaries. At a more detailed level, there are a number of interesting differences. First of all, the vowel boundary in the perception data (Figure 8B) is located at higher $F3$ than predicted. This may be an (uninteresting) effect of the specific parameter settings in the stimulus generation. $F2$ in the vowel was set at a fixed value (1800 Hz) somewhere between typical values for /i/ and /y/. If a slightly higher $F2$ had been used, listeners would have been biased somewhat toward /i/, thus shifting the vowel boundary to lower $F3$. Alternatively, the discrepancy may be a range effect. Briefly, the term *range effect* is used to describe the finding that the location of a category boundary measured in a categorization experiment depends on the stimulus range that is used in the experiment. Participants have a tendency to place the category boundary away from the endpoints (i.e., toward the middle) of a continuum (Parducci, 1965). Range effects have been shown to occur in phonetic perception and are generally larger for vowel categorization than for consonant categorization (Repp & Liberman, 1987; Rosen, 1979).

Another difference concerns the relative weighting of the two cues (F_{fr} and $F3$) in the predicted and observed categorizations. Although predicted and observed weightings are almost identical

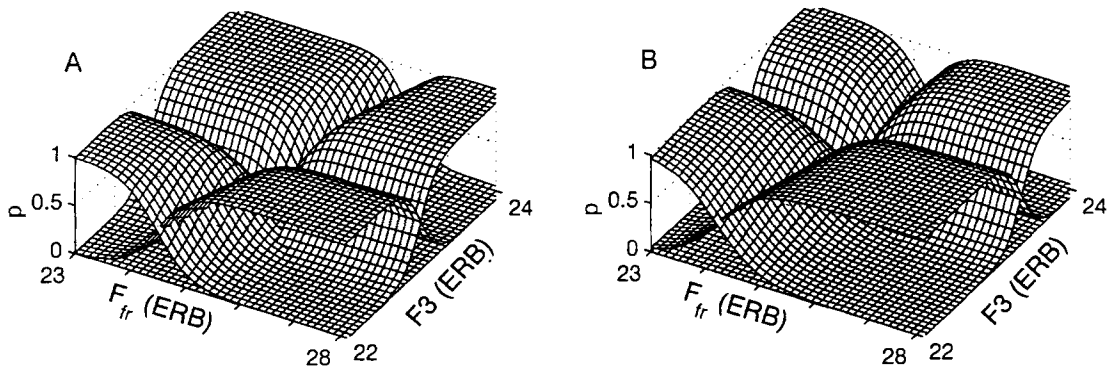

Figure 8. Comparison of the predicted (A) and observed (B) probability surfaces. The territorial plots associated with A and B are given in Figures 6D and 7, respectively. ERB = equivalent rectangular bandwidth.

for the vowel categorization (both nearly only depend on *F*3), the fricative categorization was predicted to depend more heavily on *F*3 than it actually does. This is visually evident from the fact that the fricative boundary tends more toward a diagonal orientation in $F_{fr} \times F3$ space in Figure 8A than in 8B (see also Figures 6D and 7). Again, range effects in the perception experiment may be responsible. To my knowledge, range effects causing boundary shifts have only been documented for one-dimensional stimulus continua. It seems likely that manipulation of the range in a two-dimensional continuum may also produce boundary shifts. It is conceivable, however, that the *orientation* of a category boundary may also be changed by a certain two-dimensional range manipulation. To explain this, compare 2 two-dimensional continua: one that describes a rectangle in stimulus space (a classical orthogonal design, as in Experiment 2) and one that describes a parallelogram (as, e.g., used by Sussman, Fruchter, Hilbert, & Sirosh, 1998). If listeners tend to place category boundaries away from the edges of a continuum, the boundaries will generally tend to rotate toward an orientation parallel to the sides of the continuum. Consequently, when listeners are presented with the parallelogram-shaped continuum, category boundaries are expected to be more slanted than when listeners are presented with the rectangular continuum. If the means of the natural cue distributions of four relevant categories constitute the corners of a parallelogram, whereas a rectangular continuum is used in the perception experiment (as is the case in the present experiments), the category boundaries estimated from the perception data will be biased toward an orientation parallel to the parameter axes. In the present experiment, this possibility would hardly have any effect on the vowel boundary, it being parallel as it is, but it would affect the fricative boundary in exactly the way that is observed. Of course, this explanation is only speculative at present and should be experimentally verified.

Finally, it is useful to compare the results of Experiment 2 to those of Whalen's (1989) *sushi* experiment. Recall that the HICAT analyses of Whalen's *sushi* data led to a model in which only the position of the fricative boundary depended on the vowel and in which the vowel categorization was independent of F_{fr}. First of all, the independence of the vowel categorization of F_{fr} was replicated in the analysis of the group data of the present experiment. Given that the present experiment involved a different vowel (/y/), listeners of a different language, and a larger data set, this is a striking similarity, which suggests that a more general mechanism may be operative here. A very strong hypothesis would suggest that a vowel categorization is always completely independent (acoustically as well as phonologically) of a preceding fricative or maybe even any preceding consonant. A weaker hypothesis would restrict such a claim to only some vowels or some consonants. From a pattern-recognition perspective, the findings of independence would suggest a simplification in the recognition strategy that is made possible by speakers doing their best to produce simple acoustic patterns that are easy to digest by listeners (the *orderly output constraint*; see Sussman et al., 1998; Nearey, 1997).

The dependency directions inferred from Whalen's (1989) *sushi* experiment and the present Experiment 2 are the same: The fricative categorization depends on the vowel. Recall, however, that the reliability of the direction choice for Whalen's *sushi* data was very low. The HICAT analyses suggest that different dependency types were involved in the two experiments. The best-fitting model for Whalen's data involved only position dependency, whereas for the present experiment a steepness dependency was dominant. From a pattern-classification perspective, this difference would be caused by a difference in the relevant cue distributions. Experiment 1 of the present study revealed a dominance of the convergence pattern in the acoustic data. As discussed earlier, Whalen's perception data suggest that a shift pattern is dominant in the parameters F_{fr} and *F*2 in the English syllables /si su ʃi ʃu/, which is a matter for future research.

General Discussion

The present article started with a summary of a theory of hierarchical categorization of coarticulated phonemes that is fully described elsewhere (Smits, in press). First, a theoretical analysis was presented of the possible effects of coarticulation on the distributions of acoustic cues that listeners have to deal with. Next, the HICAT model was derived representing the optimal hierarchical pattern recognition strategy given the acoustical description of the problem. The model includes three types of dependencies: dependency of the position, orientation, and steepness of one phoneme boundary on perceived value of the other phoneme.

After its conceptual and formal description, the HICAT model was applied to existing as well as new sets of phonetic categorization data with the purpose of gaining insight into processing

dependencies in the categorization of successive phonemes. First, the model was applied to data sets from two experiments published by Whalen (1989). The first of these, the *bad bet* experiment, investigated processing dependencies in the recognition of vowel and final consonant voicing in CVC words. Whalen's own analysis suggested that categorizations were not made independently. Nearey (1990) reanalyzed the same data set using an LR framework and concluded, in contrast to Whalen, that although cues were shared by the two categorizations, categorizations were made independently. A stimulus-independent diphone bias was the only transsegmental element needed in the processing model suggested by Nearey.

HICAT reanalyses of the *bad bet* experiment provided an alternative account to Nearey's (1990), one that makes a step toward the interpretation offered by Whalen (1990). In the HICAT account of the data, listeners' categorization of the final consonant depends on the perceived vowel. This is compatible with Whalen's conclusion of a categorization dependency, although the present analyses support a one-way dependency rather than the bidirectional dependency favored by Whalen (1989).

Nearey's (1990) analysis of Whalen's (1989) *sushi* data provided further support for his hypothesis of independent phoneme recognizers plus a diphone bias. A HICAT reanalysis showed that a hierarchical categorization hypothesis gives an equally good account of the *sushi* data. Although the analysis provided evidence that the fricative categorization was dependent on the perceived vowel, this conclusion was unreliable because the data set was small.

Comparisons of the performances of Nearey's (1990) LR model and of HICAT presented at various points in the article have made it clear that it is not a trivial matter to decide between the information-processing assumptions underlying the two models. Are phoneme categorizations truly independent, with only (stimulus-independent) diphone biases operating at higher levels of processing, or are there hierarchical processing dependencies involved in the categorization of successive phonemes? On a theoretical level, it could be argued that although the two models give comparable performance in terms of GOF, Nearey's account of the data is more parsimonious, and therefore preferable, because it does not include categorization dependencies. This conclusion does not hold for two reasons. First of all, compare Nearey's diphone-biased LR model to the most similar version of the HICAT model, that is, HICAT including only a boundary position dependency. On the small scale of the *bad bet* experiment, the complexity of the two models is equal because they have equal numbers of parameters. On a life-size scale, that is, in a real speech recognition system able to deal with all sounds of a language, the two models have the same boundaries on their complexity: The maximum number of extra diphone parameters needed in the LR model and the maximum number of extra c_0 parameters in HICAT are both equal to the number of diphones in the language. Second, although the finally selected HICAT model for the *bad bet* data does include one more parameter than Nearey's, the extra parameter (c_β) is responsible for a significant increase in GOF. The model without this parameter performs as well as Nearey's. The fact that Nearey has not found a similar extra processing dependency is related to the mathematical form of the LR model. There is no equivalent of the hierarchical steepness dependency in the LR model. A version of the LR model that would be most similar to HICAT with steepness dependency would have to include several extra parameters.

At a more practical level, that is, in terms of GOF, the analysis of Whalen's (1989) *bad bet* data as well as the present Experiment 2 suggests a small advantage for the hierarchical account over the independent account. However, so far only a few cases have been examined, and the differences in performance between the two models are small. Therefore, it is too early to decide between the two accounts of processing of coarticulated phonemes. Unfortunately, there is at present no satisfactory method of deciding between the two models apart from simply comparing levels of GOF. One aspect of the process that has so far been neglected is the time course of the categorization of two successive phonemes. Studying this aspect may suggest which of the two models is more appropriate. This is a topic for future research.

After the reanalyses of Whalen's (1989) data, two new experiments were presented. The first experiment addressed the question of what the training material for the phoneme recognizers looks like. One important cue for the fricative categorization (F_{fr}) and one for the vowel ($F3$) were measured on a set of syllables /si sy ʃi ʃy/, spoken by 17 male talkers. The resulting cue distributions showed that the fricative parameter F_{fr} was highly dependent on the vowel context and displayed both a shift and a convergence pattern. The distributions of $F3$, on the other hand, were more or less independent of the fricative context. Quantitative predictions showed that a pattern recognizer would benefit most from introducing a hierarchical dependency in which the fricative categorization depends on the vowel. Among the three dependency types, the steepness dependency was predicted to be most important.

These predictions were tested in Experiment 2. A two-dimensional continuum was created by generating synthetic stimuli in which analogues of the acoustic parameters F_{fr} and $F3$, measured in Experiment 1, were orthogonally varied. Listeners' categorizations of these stimuli were analyzed with the HICAT model. The analysis of the pooled data revealed that (a) there was a significant hierarchical dependency; (b) the fricative categorization depended on the perceived vowel; (c) the steepness dependency was the dominant dependency type; (d) the vowel categorization was independent of F_{fr}. All these findings are in perfect agreement with the predictions made on the basis of the acoustic measurements of Experiment 2. Furthermore, a qualitative comparison of predicted and observed locations, orientations, and steepnesses of category boundaries showed good agreement. However, analyses of individual participant data revealed substantial variation, with one of the participants showing evidence of a categorization hierarchy opposite to the one found for the group data. This variability makes the above conclusions based on the group data somewhat tentative.

It is generally accepted that each phonetic contrast is acoustically multidimensional and that listeners perceiving the contrast are sensitive to changes in many acoustic properties (e.g., Diehl & Kluender, 1987). The question therefore arises whether the predictions and observations of processing dependencies presented in this study, being based on measurement and variation of only two acoustic properties for two phonetic contrasts, can be generalized to everyday speech perception.

First of all, it is important to realize that several studies before the present one have (successfully) made predictions about phonetic categorization on the basis of acoustic measurements of low dimensionality (e.g., Nearey & Hogan, 1986). The present study is only first in making predictions about *dependencies* in phonetic categorization. Indeed, because it is merely a first step toward solving the problem, it seemed sensible to start as simple as possible, using a two-dimensional acoustic space. Of course, the question remains whether I would have arrived at a different prediction if I had measured different acoustic cues. Of this I cannot be entirely certain. It is, however, important to reiterate that the pattern of acoustic cue distributions that I found was produced by an assimilation process. Because rounding spreads from the vowel to the fricative, the fricative spectrum is strongly affected by the vowel, whereas the vowel spectrum is hardly affected by the fricative (see Figures 4 and 5). It is therefore not unreasonable to expect similar patterns of distributions, and therefore similar predictions, to hold for other acoustic cues.

The use of a two-dimensional stimulus continuum to test the predictions is similarly not unprecedented. Several past studies have investigated potential dependencies in phonetic categorization using two-dimensional stimulus continua (Massaro & Cohen, 1983; Whalen, 1989). As long as the continuum evokes convincing phonetic contrasts, which I demonstrated to be the case for Experiment 2, it is reasonable to expect the experimentally obtained dependencies to generalize to the multidimensional case.

The research presented in this article is based on the assumption that listeners categorizing coarticulated phonemes essentially behave like statistical pattern recognizers. During their lives, adult listeners have been presented with a large set of acoustic patterns. They have estimated statistical properties of these patterns, and they have selected appropriate strategies to classify them. In more specific terms relevant to the present experiments, listeners have stored statistical descriptions of acoustic cues relevant to the recognition of fricatives and vowels in memory and have potentially settled on certain hierarchical dependencies to increase performance in the relatively difficult case of feature assimilation.

The present research represents a strong test of the basic assumption. First of all, like the studies by Whalen (1989) and Nearey (1990, 1997), the present experiments and model analyses tried to uncover processing dependencies in the listeners' categorization of coarticulated phonemes. In addition, however, I attempted to provide a detailed answer to why the recognizer may benefit from the dependencies and to actually predict the dependencies used by listeners on the basis of a set of acoustic measurements. This attempt was partly successful. On the one hand, the analysis of the pooled categorization data was in close agreement with the predictions. The variability in the individual analyses, on the other hand, warrants some caution in drawing strong conclusions about the role of hierarchical strategies in phonetic categorization. Future research will have to focus on reliable methods to decide between the two dependency directions and between the HICAT model and competing categorization models.

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.

Batchelder, W. H., & Crowther, C. S. (1997). Multinomial processing tree models of factorial categorization. *Journal of Mathematical Psychology, 41,* 45–55.

Boersma, P., & Weenink, D. (1999). *Praat, a system for doing phonetics by computer* [On-line manual]. Retrieved from: http://www.fon.hum.uva.nl/praat/.

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

Collins, B., & Mees, I. (1981). *The sounds of English and Dutch.* The Hague, the Netherlands: Leiden University Press.

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

Fowler, C. A. (1984). Segmentation of coarticulated speech in perception. *Perception & Psychophysics, 36,* 359–368.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition.* New York: Academic Press.

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

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

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics, 28,* 407–412.

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]–[s] distinction. *Perception & Psychophysics, 28,* 213–228.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle.* Cambridge, MA: MIT Press.

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

Massaro, D. W., & Oden, G. C. (1980). Evaluation and integration of acoustic features in speech perception. *Journal of the Acoustical Society of America, 67,* 996–1013.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear modeling.* 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.

Nearey, T. M. (1990). The segment as a unit of speech perception. *Journal of Phonetics, 18,* 347–373.

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

Nearey, T. M. (in press). On the factorability of phonological units in speech perception. In *Papers in laboratory phonology VI.* Cambridge, UK: Cambridge University Press.

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.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72,* 407–418.

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

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

Rosen, S. M. (1979). Range and frequency effects in consonant categorization. *Journal of Phonetics, 7,* 393–402.

Schatz, C. D. (1954). The role of context in the perception of stops. *Language, 30,* 47–56.

Scheffers, M. T. M. (1983). *Sifting vowels.* Unpublished doctoral thesis, Groningen University, the Netherlands.

Smits, R. (1997). A pattern-recognition-based framework for research on phonetic perception. In *Speech hearing and language, work in progress 9* (pp. 195–229). London: Department of Phonetics and Linguistics, University College London.

Smits, R. (in press). Hierarchical categorization of coarticulated phonemes: A theoretical analysis. *Perception & Psychophysics.*

Stevens, K. N. (1998). *Acoustic phonetics.* Cambridge, MA: MIT Press.

Sussman, H. M., Fruchter, D., Hilbert, J., & Sirosh, J. (1998). Linear correlates in the speech signal: The orderly output constraint. *Behavioral and Brain Sciences, 21,* 241–299.

Whalen, D. H. (1981). Effects of vocalic formant transitions and vowel quality on the English [s]–[ʃ] boundary. *Journal of the Acoustical Society of America, 69,* 275–282.

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

Whalen, D. H. (1992). Perception of overlapping segments: Thoughts on Nearey's model. *Journal of Phonetics, 20,* 493–496.

Received September 8, 1999
Revision received November 29, 2000
Accepted December 15, 2000 ■