Perception & Psychophysics
2001, 63 (7), 1109-1139

Hierarchical categorization of coarticulated phonemes: A theoretical analysis

ROEL SMITS
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

This article is concerned with the question of how listeners recognize coarticulated phonemes. The problem is approached from a pattern classification perspective. First, the potential acoustical effects of coarticulation are defined in terms of the patterns that form the input to a classifier. Next, a categorization model called HICAT is introduced that incorporates hierarchical dependencies to optimally deal with this input. The model allows the position, orientation, and steepness of one phoneme boundary to depend on the perceived value of a neighboring phoneme. It is argued that, if listeners do behave like statistical pattern recognizers, they may use the categorization strategies incorporated in the model. The HICAT model is compared with existing categorization models, among which are the fuzzy-logical model of perception and Nearey's diphone-biased secondary-cue model. Finally, a method is presented by which categorization strategies that are likely to be used by listeners can be predicted from distributions of acoustical cues as they occur in natural speech.

How do listeners decode coarticulated speech? Throughout the history of speech perception research, this has been one of the central issues. There are two essential aspects to the issue that cannot be addressed independently. The first is the question of the basic recognition unit: What is the size of the unit onto which the acoustic signal is mapped initially? After some 50-odd years of speech perception research, there is no consensus on this issue. There are still active supporters for units across the entire range of possible sizes, such as the distinctive feature (Lahiri & Reetz, 1999; Stevens, 1995), the phoneme (Nearey, 1990, 1997; Norris, McQueen, & Cutler, 2000), the syllable (Massaro, 1998; Segui, Frauenfelder, & Mehler, 1981), and the word (Goldinger, 1998; Johnson, 1997).

The second aspect of the issue of how listeners decode coarticulated speech is concerned with the recognition process itself. Owing to coarticulation, the acoustic realization of any linguistic symbol in the speech stream is affected by surrounding symbols, irrespective of the size of the recognition unit. Does listeners' categorization of linguistic symbols reflect these dependencies and, if so, in what way? In other words, what are the processing dependencies, if any, in the recognition of linguistic symbols?

Part of this research was carried out at the Department of Phonetics and Linguistics, University College London. This part was supported by a NATO-Science Fellowship and by a TMR Fellowship granted by the European Commission. The author is grateful to Terry Nearey, Louis ten Bosch, Stuart Rosen, James McQueen, and Anne Cutler for help and encouragement and to John Kingston, Gregg Oden, and an anonymous reviewer for comments on earlier versions of the paper. Correspondence concerning this article should be addressed to R. Smits, Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands (e-mail: roel.smits@mpi.nl).

The present research approaches the two issues introduced above from a pattern classification perspective. First, the problem is redefined in terms of the acoustic patterns that form the input to a classifier. Next, a categorization model is introduced that incorporates hierarchical dependencies to optimally deal with this input. If listeners do behave like statistical pattern recognizers, they may use strategies incorporated in the model. Finally, a method is presented for predicting categorization strategies that are likely to be used by listeners on the basis of distributions of acoustical cues as they occur in natural speech.

Testing for processing dependencies in phonetic categorization is far from straightforward, as is evidenced by the lack of consistency of the findings reported in the literature. For example, Mann (1980), Whalen (1989), and Wood and Day (1975) claim to have found dependencies between the recognition of successive phonemes, whereas Fletcher (1953), Massaro and Cohen (1983), and Nearey (1997) claim independence. These apparent discrepancies in the reported findings may result from the fact that the different experimental methodologies adopted in the various studies actually tapped into different components of the phonetic categorization process. The process that maps the acoustic signal onto linguistic units is generally assumed to consist of several distinct processing stages, such as auditory stimulus encoding, extraction of relevant acoustic cues, mapping of the cue vector onto response probabilities, and response selection (e.g., Massaro, 1987; Smits, 1997). Distinct types of processing dependencies may be associated with each of these processing stages, which may have been confounded in some of the studies mentioned above.

History has shown that two types of dependency are particularly prone to confounding—namely, acoustic de-

pendency and phonological dependency. There exists an acoustic dependency between two phonetic categorizations if both share acoustic cues. Consider the example of the recognition of the liquid and stop in nonsense words ALDA ARDA ALGA ARGA, as first studied by Mann (1980). If the third formant frequency at the offset of the liquid is used in both the stop classification and the liquid classification, there is an acoustic dependency between the two. If, on the other hand, the stop classification depends on the perceived liquid *identity*, there is a phonological dependency. Mann's experiments showed that at least one of the two types of dependencies was present, but they could not distinguish between the two. Recently, Lotto and Kluender (1998) showed that at least part of the observed boundary shift in the stop categorization has an acoustic/auditory basis.

Relatively recently, a number of researchers have advocated a pattern recognition approach to the study of phonetic perception (Massaro, 1987, 1998; Nearey, 1990, 1997; Smits, 1997). The pattern recognition approach has two essential components. First, it is assumed that many aspects of the phonetic categorization process in adult listeners are the result of *training* on previously perceived and categorized speech. In other words, at least some of the strategies and parameter settings of the system are based on acoustical statistics of natural speech. Second, the approach advocates quantitative modeling of the entire recognition process to unravel certain aspects of the process of phonetic categorization, such as the basic recognition unit and processing dependencies. By explicitly modeling the various levels of information processing and their possible interactions, the previously mentioned confounding between acoustic and phonological context effects may be avoided.

Within the pattern recognition framework, both Massaro and colleagues (Massaro, 1987, 1998; Oden & Massaro, 1978) and Nearey (1990, 1997) have defined models for the categorization of coarticulated speech (these models will be discussed extensively in later sections). The present article introduces a new model, called HICAT (from hierarchical categorization), which represents a competitor to Massaro's and Nearey's models. The core distinguishing feature of the new model is that it incorporates a particular type of processing dependency—that is, *hierarchical* dependency, which has been frequently hypothesized in the literature but never explicitly captured in a quantitative model (e.g., Carden, Levitt, Jusczyk, & Walley, 1981; Eimas, Tartter, Miller, & Keuthen, 1978; Harris, 1958; Miller, 1981; Nearey, 1992). A *hierarchical categorization* can be defined as the case in which one categorization depends on the output of another categorization, but not vice versa. In other words, there is a one-way phonological dependency. Neither Massaro's nor Nearey's model incorporates such hierarchical dependencies.

In the formulation of the HICAT model, the relation between the effects of coarticulation on the acoustic manifestation of successive phonemes and the usefulness and nature of potential processing dependencies in the human speech recognition system plays a central role. The explicit relation between natural cue distributions and phonetic categorization has so far not received much attention in Massaro's work, whereas in the research of Nearey and colleagues, quantitative predictions of categorization behavior from acoustic data have been limited to single-phoneme categorizations (Andruski & Nearey, 1992; Hillenbrand & Nearey, 1999; Nearey & Assman, 1986; Nearey & Hogan, 1986). As will be shown in later sections, the HICAT model is set up so that it produces optimal hierarchical categorization, given certain geometries in cue distributions that are likely products of coarticulation. In addition, a method is presented that predicts what types of dependencies are most likely to occur in listeners' categorization strategies, given a set of acoustic cue distributions.

The paper is structured as follows. In the next section, a theoretical consideration is presented of the potential effects of coarticulation on the statistical distributions of acoustic cues. An examination of how a pattern classifier would deal with this problem reveals that, under certain conditions, classification performance would benefit from the use of certain types of hierarchical categorization dependencies. Next, the HICAT model for the hierarchical categorization of phonemes that encorporates such dependencies is presented. HICAT is compared with competing categorization models—namely, the fuzzy-logical model of perception (FLMP), the diphone-biased secondary-cue model (DBSCM), and general recognition theory (GRT). The following section gives a simple formula by which the reliability of the inference of dependency direction can be estimated for practical HICAT model fits. Finally, a technique is presented for predicting processing dependencies from acoustical measurements on natural utterances.

PATTERN RECOGNITION OF COARTICULATED PHONEMES

Acoustic Consequences of Coarticulation

How does the pattern recognition approach to phonetic perception address the problem of the decoding of coarticulated phonemes? Recall that an essential assumption of the approach is that listeners base their recognition strategies on statistics of relevant acoustic parameters. Nearey (1992) considered the recognition of the English syllables /si su ʃi ʃu/ from a pattern classification angle. On the basis of the mean values of the frequencies of the second formant and the main fricative resonance (the dominant cues in this recognition problem) derived from Soli (1981), Nearey discussed several classification strategies by which to approach optimal performance while keeping the classifier relatively simple. Below, I will try to extend Nearey's (1992) reflections in several ways, by first considering the likely consequences of coarticulation on statistical distributions of acoustic cues as they occur in natural speech, followed by an analysis of how hierarchical strategies can be employed to deal with these acoustic patterns.

First, let us consider a simple hypothetical situation concerning the production and perception of consonant–vowel (CV) syllables, specifically the Dutch CV syllables /si sy ʃi ʃy/. This set is interesting because, in Dutch, the phoneme /ʃ/ is unrounded, so the pair /s ʃ/ differ only in place of articulation, whereas the vowel pair /i y/ differ only in rounding. In the natural production of these syllables, rounding spreads from the vowel to the preceding fricative. Let us imagine, however, that this spreading of rounding did not take place and that the production of the Dutch CV syllables /si sy ʃi ʃy/ were completely defined by the following idealized characteristics: (1) The CVs are produced without coarticulation; (2) /s/ and /ʃ/ differ only in the frequency of a spectral prominence F_{fr} in the frication noise; (3) /i/ and /y/ differ only in the frequency of the third formant; (4) different tokens of the same CV syllable are distributed as a two-dimensional Gaussian stochastic variable; (5) the Gaussian distributions associated with different CV syllables have equal diagonal covariance matrices—that is, they differ only in their means. Figure 1A gives a graphic representation of this situation. Essentially, the means for /si sy ʃi ʃy/ form the corners of a rectangle.

Next, consider the case in which assumptions 1 and 2 are relaxed—that is, rounding does spread to the fricative, whereas the influence of the fricative on the vowel is negligible. Owing to assimilation, the main fricative cue F_{fr} will change with vowel context (e.g., Soli, 1981). Such change can be decomposed into two components. First, the spectral prominence in the frication noises of /sy/ and /ʃy/ may shift down in frequency by an equal amount (see Figure 1B). Second, the spectral prominences in the frication noises of /sy/ and /ʃy/ shift in opposite directions—that is, they converge (Figure 1C). In a more realistic situation, a combination of these two effects is expected to occur, because rounding generally does have the effect of lowering spectral prominences, but not by equal amounts for /s/ and /ʃ/. The reason for this is that, because the length of the front cavity (the cavity "downstream" from the fricative constriction) is smaller for /s/ than for /ʃ/, the *relative* increase in the length of the front cavity is larger for /s/ than for /ʃ/ and, therefore, the downward shift of the spectral prominence is expected to be larger for /s/ than for /ʃ/. This case is illustrated in Figure 1D. Although often a mixture of the shift and convergence effects is expected, it is useful to keep them separate, as will become clear in later sections.

The acoustic effects of coarticulation shown in Figure 1D are expected to be quite common, even in situations in which there is no actual feature spreading, as in the fricative–vowel case. Consider, for example, the much-discussed dependency of the frequency of the second formant at voice onset ($F2_o$) and in the vowel ($F2_v$) in stop-vowel syllables. Sussman and colleagues (e.g., Sussman, McCaffrey, & Matthews, 1991), and others before them, have shown that, owing to coarticulation, $F2_o$ changes with $F2_v$, but more for /b/ than for /d/. This dependency appears to be quite regular and can be well described by the so-called *locus equations*. Plotting $F2_o$ horizontally and $F2_v$ vertically, measured on several tokens of the syllables /bi/, /di/, /bu/, and /du/, will probably produce a geometry which, like Figure 1D, combines the shift and the convergence patterns. The major difference with Figure 1D would be that now the top two distributions (/bi/ and /di/) will be closer together than the bottom two (/bu/ and /du/).

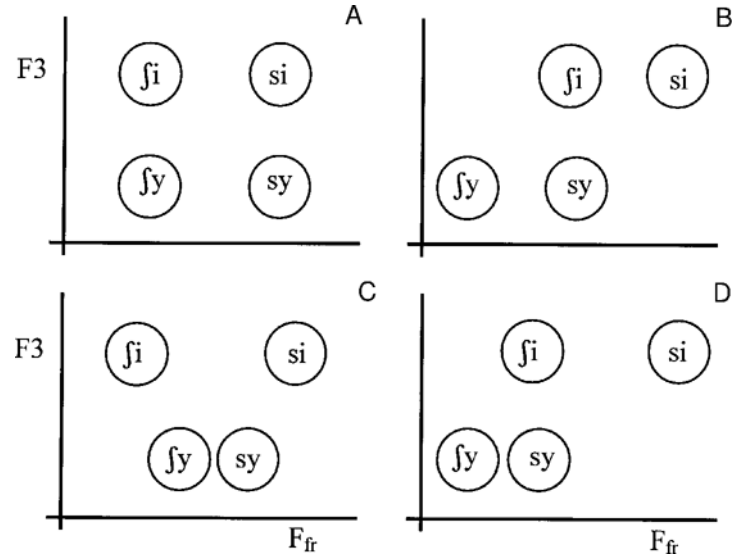

Figure 1. Geometries of distributions of acoustic cues *F3* and F_{fr} for four hypothetical varieties of coarticulation. Circles represent isoprobability contours of two-dimensional Gaussian pdfs. Panel A represents absence of coarticulation, panels B and C represent *shifted* and *converged* geometries due to coarticulation, respectively, and panel D represents a combination of B and C.

Categorization Strategies

How would a pattern recognizer deal with the situations illustrated in Figure 1? Before this question can be answered, *performance* of a classifier must be defined. As in Nearey's normal a posteriori probability model (NAPP, Nearey & Hogan, 1986), it is assumed that listeners represent phoneme categories as multidimensional Gaussian probability density functions (pdfs) that reflect the statistics of relevant acoustical cues in natural speech. Furthermore, it is assumed that listeners map an incoming sound not onto a single discrete phonological label, but instead onto phoneme *category goodness* or activation levels by reading off the likelihoods of the various phoneme pdfs, given the cue vector for the incoming sound. During normal speech recognition, these *fuzzy* goodness levels are used to activate lexical items, and no categorical decisions about phonemes are made. Only in phonetic categorization experiments, when listeners are forced to make categorical decisions at the phoneme level, are the goodness levels actually mapped onto a single phonetic response. Given this set of assumptions, the performance of a classifier is (qualitatively) defined as the similarity of the classifier's fuzzy output to the pdfs of the relevant acoustical cues.

Like the studies by Massaro and Cohen (1983), Whalen (1989), and Nearey (1990, 1992), the present research focuses on four-response categorization (4RC) experiments involving two binary decisions on two consecutive phonemes (/si sy ʃi ʃy/). In general, the optimal classifier—that is, the classifier with the highest performance—in a four-way categorization bases its categorization on the acoustical distributions associated with each of the four responses. This means that, for the present experiments, the optimal classifier uses the syllable or diphone as a unit. However, when we scale this discussion up to more realistic proportions, it is more economical to use the phoneme as a recognition unit than the syllable, because in all languages the set of phonemes is much smaller than the set of syllables or diphones. For most languages, the number of phonemes lies, roughly, between 10 and 100, whereas the number of diphones or syllables lies between 1,000 and 100,000 (Maddieson, 1984; for more elaborate considerations on the implications, see Dupoux, 1993; Nearey, in press). Therefore, it is easier to *train* a phoneme-based recognizer than a syllable-based recognizer, because it contains fewer parameters. Also, recognition is easier and faster, because the number of comparisons of the input to the categories is much smaller.

On the other hand, provided there is unlimited training material and processing capacity, a syllable-based recognizer will generally have higher performance than a phoneme-based recognizer. The question is, does this potential difference in performance warrant the use of the syllable as a unit, even at a cost of much higher complexity? Or, to put it more succinctly, has the human speech recognition system selected the quick-and-dirty option or for the slow-and-accurate option? The present research favors the quick-and-dirty option. Moreover, it will be shown that this option need not, in fact, be so dirty. Below, it will be argued that, if the acoustic consequences of coarticulation of two consecutive phonemes are so severe that phoneme-based recognition performance drops significantly, this may be repaired by introducing *dependencies* between the categorizations of the phonemes that reflect the regularities in the acoustic effects of coarticulation. Thus, the performance of a phoneme-based recognizer can be increased at the cost of a slight increase in complexity, but without switching to the syllable unit.

Given the choice of the phoneme as the unit of recognition, let us—for the moment, only qualitatively—consider likely strategies dealing with the problems illustrated in Figure 1. The pattern recognition problem of Figure 1A is easily, and optimally, solved by using a horizontal boundary separating the vowels /i/ and /y/ and a vertical boundary separating the fricatives /s/ and /ʃ/. This strategy, which is equivalent to the strategy of Figure 2a in Nearey (1992), is illustrated in Figure 2A.

The problem of Figure 1B, where the pdfs for /sy/ and /ʃy/ are shifted with respect to those of /si/ and /ʃi/, can be approached in two ways. The first is to make the fricative categorization dependent on $F3$. In Nearey's (1990, 1992, 1997) terms, the "secondary cue" $F3$ is used beside the "primary cue" F_{fr}. In the context of the present model, I will use the term *acoustic context* to indicate this situation. That is, in the categorization of a phoneme (/s/–/ʃ/), an acoustic parameter is used that is measured from the following vocalic portion. This strategy results in the nonvertical boundary between /s/ and /ʃ/, as is indicated in Figure 2B (see also Figure 2D in Nearey, 1992).

Figure 2C illustrates a different strategy for dealing with the same acoustic distributions. In this case, the fricative categorization does not depend on $F3$ in the vowel, but on the vowel *label*. The /s/–/ʃ/ boundary in context /y/ is shifted with respect to context /i/, reflecting the shift of the pdfs. This is a hierarchical dependency, because the fricative categorization depends on the vowel categorization, and not vice versa. As in Figure 2B, there is a context effect, but rather than being acoustic, it is *phonological*. Figure 2B in Nearey (1992) represents essentially the same strategy.

The distinction between acoustic and phonological context effects is essential, but, as was discussed earlier, the two are easily confused on the basis of categorization data. Suppose one were to investigate the influence of the following vowel on the categorization of a preceding fricative using an /s/–/ʃ/ continuum followed by /i/ in one condition and /y/ in another. This is comparable to making two horizontal cross-sections of the two-dimensional acoustical/perceptual space given in Figure 2. The results of such an experiment could not be used to distinguish between the strategies represented by Figures 2B (acoustical context) and by 2C (phonological context). In both cases, the categorization functions would display a shift of the category boundary.

How would a pattern classifier choose between acoustical and phonological context strategies? When

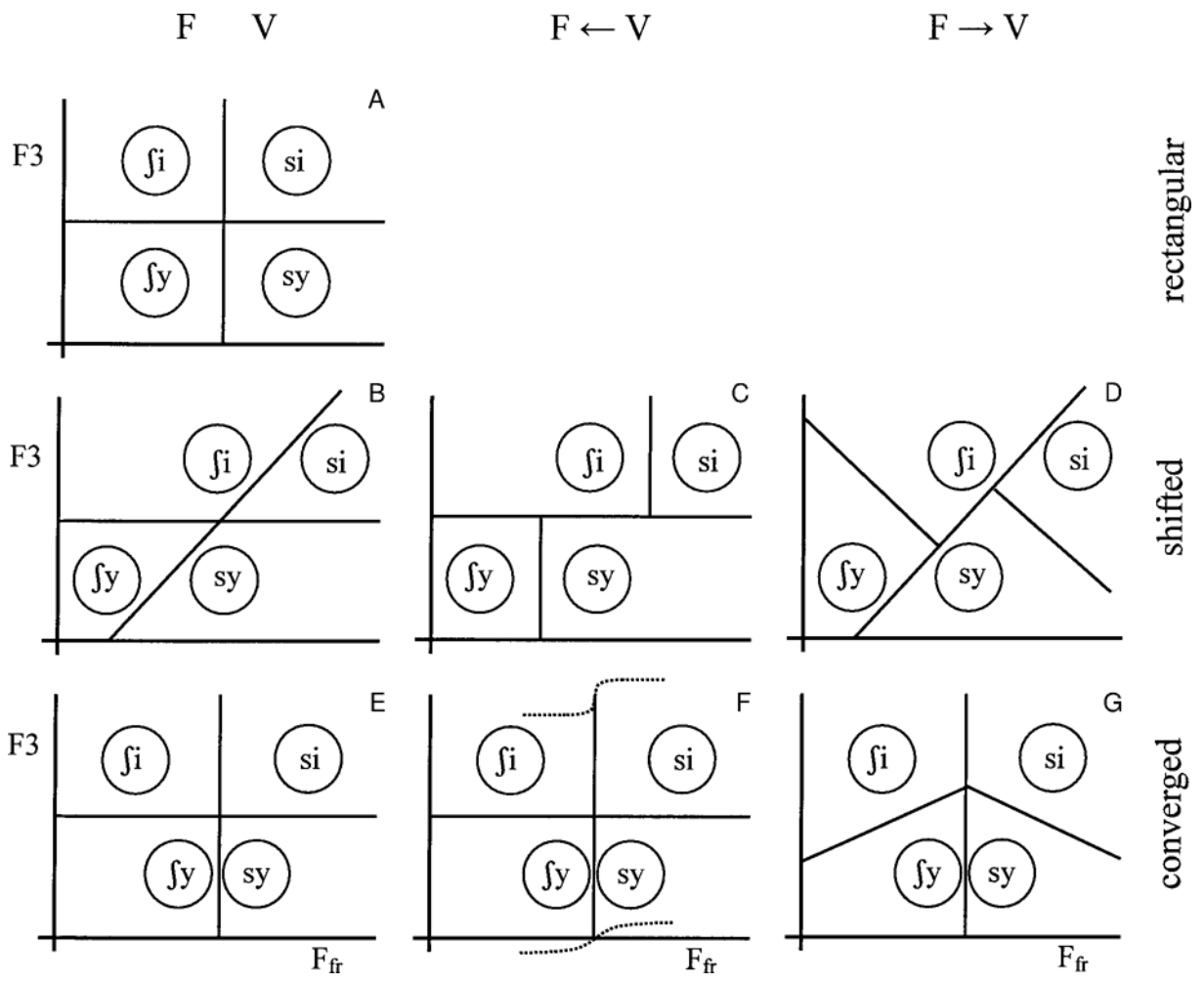

Figure 2. Categorization strategies for the pdf-geometries of Figures 1A–1C. The panels in the left-hand column (indicated by "F V") represent category boundaries for categorization strategies without use of phonological context for the rectangular (A), shifted (B), and converged configurations (E). Panels in the middle column (indicated by "F ← V") give boundaries for strategies in which the fricative categorization depends on the perceived vowel for the shifted (C) and converged (F) geometries. The dotted sigmoid shapes in panel F indicate that the /s/–/ʃ/ boundary is steeper in /i/ context than in /y/ context. Finally, the panels in the right-hand column illustrate boundaries for the reverse hierarchy in which the vowel categorization depends on the perceived fricative for the shifted (D) and converged (G) geometries.

coarticulation is moderate—that is, when the horizontal shift of the pdfs for /sy/ and /ʃy/ with respect to /si/ and /ʃi/ is small—use of acoustic context will still give good performance. When coarticulation is high, however—that is, in case of a large shift—performance seriously degrades, and the use of a hierarchical dependency becomes necessary. (Note that the circles in Figures 1 and 2 indicate isoprobability contours, not the ranges, of the cue distributions—that is, it is assumed that there is significant overlap between the four distributions.) If a strategy involving phonological context is chosen, how to decide which categorization should depend on which is not a trivial matter. Figure 2D represents a hierarchy opposite to that of Figure 2C—that is, the vowel categorization now depends on the fricative categorization. As in Figure 2C, this strategy results in better performance than does the acoustic context strategy (Figure 2B). Possible ways of choosing the dependency direction will be discussed later.

At a first glance, the categorization problem of Figure 1C is no more difficult than that of Figure 1A. A simple horizontal and vertical boundary would give good separation. However, as will be explained later in the quantitative section, the performance of a classifier that bases its label selection on *fuzzy match* of the input to the pdfs for the relevant categories can be increased by the introduction of a hierarchical dependency. Again, both dependency directions are possible. If the fricative categorization depends on the vowel categorization, the category boundaries are indeed horizontal and vertical, as in Figure 2A, but there is a dependency in terms of the steepness of the fricative categorization functions on the

vowel. Because the fricative pdfs in /i/ context are well separated, the /si/–/ʃi/ categorization would produce a steep categorization function. The fricative pdfs in context /y/, on the other hand, are less well separated, which results in a shallow /sy/–/ʃy/ categorization function. This type of hierarchical dependency is illustrated in Figure 2F, where the dotted sigmoid functions below and above indicate the boundary steepness below and above the horizontal boundary. If the vowel categorization depends on the fricative categorization, the recognition performance can be improved, as compared with the strategy of Figure 2E, by adjusting the orientations of the /si/–/sy/ boundary and the /ʃi/–/ʃy/ boundary in opposite directions, as is illustrated in Figure 2G. Note that the latter two types of hierarchical categorization were not considered by Nearey (1992), who only discussed boundary-shift dependence.

In summary, when neighboring phonemes are severely coarticulated, a pattern recognizer that makes independent phoneme categorizations may perform less than optimally. Introducing a hierarchical dependency between the categorizations—that is, making one of the phoneme categorizations dependent on the other—may repair this drop in performance. Different patterns in the acoustic effects of coarticulation call for different types of dependency—that is, (1) dependency of the boundary *location* (Figures 2B and 2C), (2) dependency of the boundary *steepness* (Figure 2F), or (3) dependency of the boundary *orientation* (Figure 2G).

Earlier, it was mentioned that, from a pattern classifier's point of view, hierarchical categorization dependencies are particularly beneficial in cases in which assimilation processes cause the geometry of cue distributions to deviate strongly from the rectangular pattern (Figure 1A). The following questions arise: (1) When exactly should a dependency be used? and (2) which direction should it have? These issues should be considered within the contexts of the full-scale speech recognition system and the complete phoneme inventory in a language. Instead of ad hoc procedures for each diphone, more general strategies may be used. For example, a possible general strategy might be to make Phoneme 1 dependent on Phoneme 2 in cases in which there is feature assimilation from 2 to 1. In the present case, the recognition of a fricative would then be made dependent on the first vowel following the fricative. Alternatively, dependencies may simply be associated with all possible phoneme pairs, where some dependencies will be strong and experimentally detectable, whereas others will be very weak and difficult to observe. In this first theoretical study, these issues are left open.

THE HICAT MODEL

Mathematical Definition

For explanatory reasons, I will take a shortcut in the presentation of the mathematics of the HICAT model. The model defined in terms of the response probabilities as a function of the stimulus coordinates is remarkably simple and interpretable. The three types of hierarchical dependencies introduced earlier (dependency of boundary position, steepness, and orientation) are explicitly modeled, each represented by a single parameter. In the next section and Appendix A, I will show that HICAT has optimal conditional phoneme-based categorization performance, given the geometries of acoustical cue distributions presented earlier in Figures 1 and 2. These derivations are somewhat more complex and might distract some readers from the important results.

The model is restricted to four-alternative forced-choice experiments, where the four alternatives arise from two orthogonal binary phoneme categorizations (as in the /si sy ʃi ʃy/ example). Also, the stimulus space is restricted to only two dimensions. It is assumed that a stimulus, represented by vector $\vec{\phi}$ in physical space Φ, is mapped onto vector $\vec{\psi}$ in psychological space Ψ by unidimensional monotonic mappings:

$$\psi_x = \mathcal{M}_x(\phi_x) \tag{1}$$

$$\psi_y = \mathcal{M}_y(\phi_y), \tag{2}$$

where $\mathcal{M}$ is an "appropriate" monotonic psychophysical mapping for physical quantity ϕ. Examples of such psychophysical mappings are the square root for auditory durations (Nearey, 1990), and the equivalent rectangular bandwidth scale for auditory frequencies (Glasberg & Moore, 1990). It is possible that the actual psychological axes are not just "warped" physical axes, as in Equations 1 and 2 but, instead, are linear combinations of warped physical axes. The HICAT model is insensitive to such linear transformations, because the ψ_x and ψ_y in turn undergo a linear transformation in Equations 3 and 4 below. Therefore, Equations 1 and 2 do not impose unnecessary restrictions.

In psychological space Ψ, we define two axes α and β as follows:

$$\alpha = p_0 + p_x\psi_x + p_y\psi_y \tag{3}$$

$$\beta = q_0 + q_x\psi_x + q_y\psi_y. \tag{4}$$

Parameter vectors $\vec{p}$ and $\vec{q}$ are chosen so that, along axes α and β, the probabilities of choosing either of the two alternatives (A_1 vs. A_2 for α, and B_1 vs. B_2 for β) change most rapidly.

Let $p(A_1 \mid S_i)$ indicate the conditional probability that stimulus S_i receives the label A_1 on categorization A. The probabilities of choosing either alternative on categorization A are assumed to be related to α by a logistic function. This is most conveniently expressed as a log-odds ratio (e.g., Bishop, Fienberg, & Holland, 1975):

$$\log\frac{p(A_1 \mid S_i)}{p(A_2 \mid S_i)} = \alpha. \tag{5}$$

The same holds for distinction B and axis β, except that here a term is added:

$$\log\frac{p(B_1 \mid A_1, S_i)}{p(B_2 \mid A_1, S_i)} = \beta + (c_0 + c_\alpha\alpha + c_\beta\beta) \tag{6}$$

$$\log\frac{p(B_1 \mid A_2, S_i)}{p(B_2 \mid A_2, S_i)} = \beta - (c_0 + c_\alpha\alpha + c_\beta\beta). \quad (7)$$

The parenthetical term in the right-hand terms of Equations 6 and 7 represents the dependency of categorization B on A. Crucially, the sign of this term depends on the outcome of the categorization A. It is noted that Equations 3 and 4 allow only a linear relationship between α and β and the psychological axes. Of course, these may, in certain cases, be extended to include quadratic or higher order terms, while preserving the hierarchical dependencies.

The categorization dependency defined by the parenthetical term has three components associated with parameters c_0, c_α, and c_β: c_0, dependency of the *position* of the B boundary on categorization A, c_α, dependency of the *orientation* of the B boundary on categorization A, and c_β, dependency of the *steepness* of the B boundary on categorization A.

The probabilities of choosing either of the four syllables are calculated simply by multiplying the appropriate (conditional) phoneme probabilities:

$$p(A_1, B_1 \mid S_i) = p(A_1 \mid S_i)p(B_1 \mid A_1, S_i) = \frac{1}{1+\exp(-\alpha)} \cdot \frac{1}{1+\exp(-\beta - c_0 - c_\alpha\alpha - c_\beta\beta)} \quad (8)$$

$$p(A_1, B_2 \mid S_i) = p(A_1 \mid S_i)p(B_2 \mid A_1, S_i) = \frac{1}{1+\exp(-\alpha)} \cdot \frac{1}{1+\exp(\beta + c_0 + c_\alpha\alpha + c_\beta\beta)} \quad (9)$$

$$p(A_2, B_1 \mid S_i) = p(A_2 \mid S_i)p(B_1 \mid A_2, S_i) = \frac{1}{1+\exp(\alpha)} \cdot \frac{1}{1+\exp(-\beta + c_0 + c_\alpha\alpha + c_\beta\beta)} \quad (10)$$

$$p(A_2, B_2 \mid S_i) = p(A_2 \mid S_i)p(B_2 \mid A_2, S_i) = \frac{1}{1+\exp(\alpha)} \cdot \frac{1}{1+\exp(\beta - c_0 - c_\alpha\alpha - c_\beta\beta)} \quad (11)$$

Note that, if $c_0 = c_\alpha = c_\beta = 0$, categorizations A and B are independent in the statistical sense, because in that case,

$$p(A_1, B_1 \mid S_i) = p(A_1 \mid S_i)p(B_1 \mid S_i), \quad (12)$$

and similarly for the other three syllables. For example, let A and B represent the /i/–/y/ and /s/–/ʃ/ distinctions, respectively. Fricative and vowel are categorized independently for a given stimulus if, for that stimulus,

$$p(/\text{si}/) = p(/\text{s}/)p(/\text{i}/), \quad (13)$$

and similarly for the other syllables.

The full dependency model defined above has nine free parameters: p_0, p_x, p_y, q_0, q_x, q_y, and the dependency parameters c_0, c_α, and c_β. The independent categorization model is nested under the dependent model and has only six free parameters (p_0, p_x, p_y, q_0, q_x, and q_y). The independent model coincides with Nearey's four-alternative logistic regression model without any diphone terms.

Appendix A contains the quantitative derivation of HICAT from the distributions of acoustical cues that were introduced in Figures 1 and 2. It is shown that the three dependency strategies associated with the boundary location, orientation, and steepness, as implemented in HICAT, give the optimal conditional categorization of the cue distributions of Figure 1. Appendix A also lists the relationships between HICAT's parameters and the parameters of the distributions of Figure 1.

Theoretical Examples

In order to illustrate the effects of the three types of hierarchical dependencies on the probability surfaces and territorial maps for the four responses, this section presents some simple theoretical examples. Figure 3 gives graphical representations of four cases: independence, position dependence, orientation dependence, and steepness dependence.

For all four cases, the relations between α and β psychological axes ψ_x and ψ_y are simple:

$$\alpha = \psi_y \quad (14)$$

$$\beta = \psi_x, \quad (15)$$

by setting $p_y = q_x = 1$ and $p_0 = p_x = q_0 = q_y = 0$. The top two panels of Figure 3 (A and B) represent independence ($c_0 = c_\alpha = c_\beta = 0$). The left panel gives the probability surfaces in psychological space; the right panel (*territorial plot*) gives only the syllable boundaries—that is, the points in psychological space at which the probabilities of choosing two adjacent syllables are equal. The A_1–A_2 boundary is a horizontal line, and the B_1–B_2 boundary is a vertical line. Note that the fact that the lines are parallel to the coordinate axes stems from the choice of vectors $\vec{p}$ and $\vec{q}$ (Equations 14 and 15) and is unrelated to phonological independence.

Panels C and D (second row) of Figure 3 display the shape of the model when c_0 is nonzero ($c_0 = 1$). Panel D shows that the B_1–B_2 boundary is shifted to the left for A_1 and to the right for A_2. Interestingly, the shift does not leave the A_1–A_2 boundary itself unaffected. As can be seen in panel D, the A_1–A_2 boundary is not linear anymore (i.e., not a straight line), except for the segment separating categories A_1B_1 and A_2B_2. The representation of a boundary position dependency, as given in Figure 2B (as well as Figure 2B in Nearey, 1992), should really be viewed within the context of a boundary-based model like GRT (Ashby & Townsend, 1986). In a similarity-based model like HICAT, territorial plots with the shape of Figure 2B can occur only when the categorization function associated with the primary categorization is a

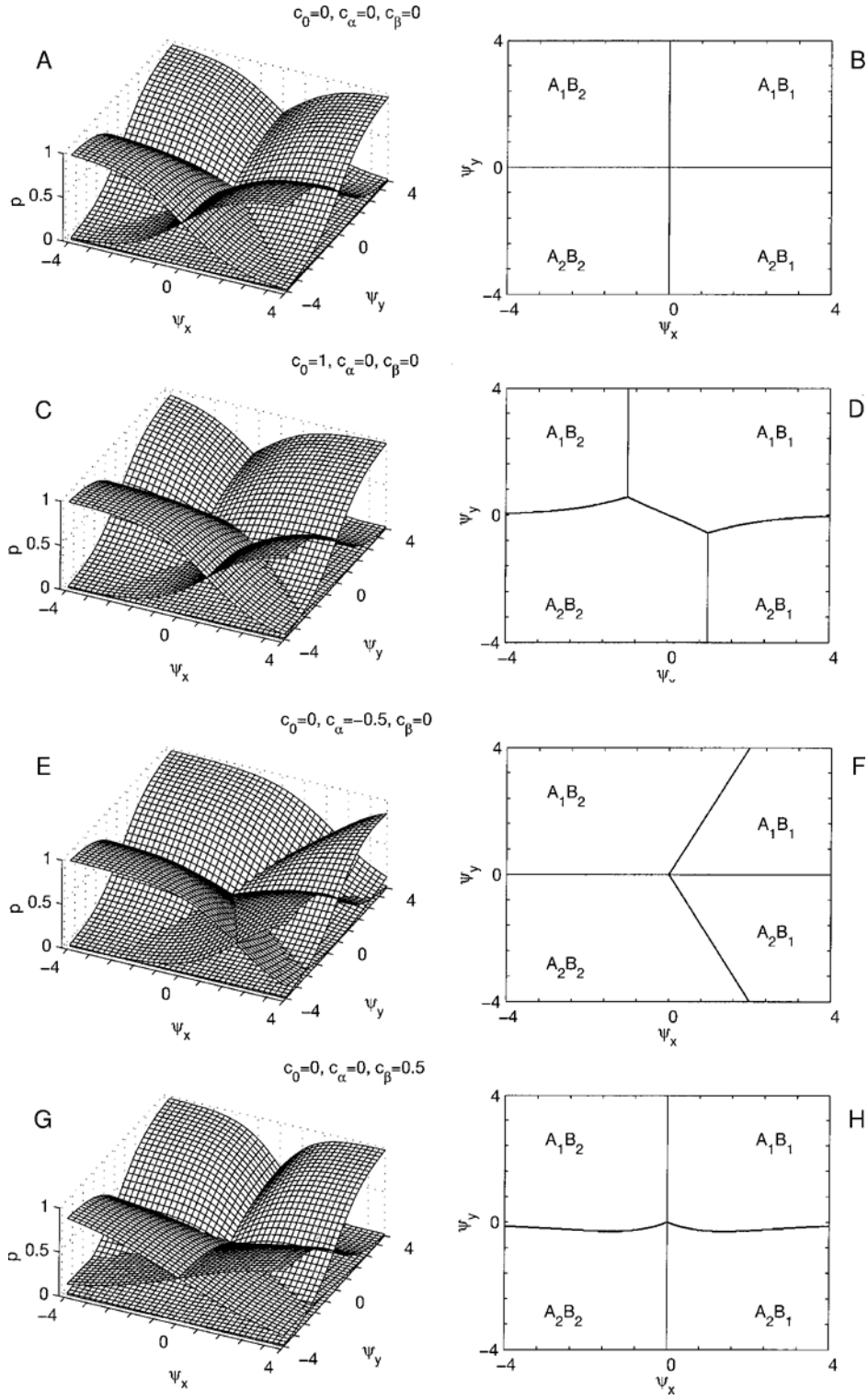

Figure 3. Illustration of the effect of three types of dependencies on response surfaces (left column) and territorial plots (right column). Panels A and B represent independence, panels C and D represent position dependence ($c_0 = 1, c_\alpha = c_\beta = 0$), panels E and F represent orientation dependence ($c_\alpha = -0.5, c_0 = c_\beta = 0$), and panels G and H represent steepness dependency ($c_\beta = 0.5, c_0 = c_\alpha = 0$). For all figures, $p_y = q_x = 1, p_0 = p_x = q_0 = q_y = 0$.

step function—that is, is entirely noise-free—which is unrealistic (see the section on model comparison, below). On the whole, the territorial plot of panel D is reminiscent of Nearey's diphone-biased primary-cue model, although they are not identical. The similarities are more fully discussed in a later section.

Panels E and F (third row) give the probability surfaces and territorial plot, respectively, for the boundary orientation dependency ($c_\alpha = -0.5$). The B_1–B_2 boundary has rotated clockwise for A_1 and counterclockwise for A_2.

Finally, panels G and H give the probability surfaces and territorial map for the steepness dependency ($c_\beta = 0.5$). Panel G shows this dependency most clearly: The two response surfaces to the "front" of the graph (for negative ψ_y) are rather shallow, whereas the two toward the "back" of the graph are much steeper. Again, the A_1–A_2 boundary is affected by this dependency. Because the probabilities for categories A_1B_1 and A_1B_2 rise more quickly with ψ_x, they "gain" somewhat on categories A_2B_1 and A_2B_2, as is evident in panel H.

Figure 3 illustrates that whenever either c_0 or c_β is nonzero, some of the boundaries shown in the territorial plot—that is, the boundaries associated with categorization A—are nonlinear. This may seem somewhat counterintuitive, given that only linear logistic functions are used in the above model definition. The nonlinearity can be understood as follows. The category boundaries for phoneme categorizations A and B, as defined in Equations 5–7, are linear, as would be evidenced by territorial plots for the phoneme categorizations. What is visible in the territorial plots of Figure 3, however, are the boundaries between *syllable* regions A_1B_1, A_2B_1, and so on, not between phoneme regions. These boundaries are derived from the syllable probabilities, which are products of (conditional) phoneme probabilities (see Equations 8–11). It is easy to evaluate from Equations 8–11 that the syllable boundaries are linear only if $c_0 = c_\beta = 0$.

Information Processing Architecture

The definition of the HICAT model given above directly links response probabilities to stimulus characteristics. An important remaining issue is the processing architecture implementing the dependency strategies. Basically, two strategies are possible: serial or parallel. The mathematical definition of HICAT seems to suggest a serial architecture. However, it will be argued below that a parallel architecture is not only compatible with HICAT's definition, it is also theoretically more appealing. The various processing steps in the HICAT model that were omitted in the earlier mathematical model definition will be fully defined in the description of both architectures.

Serial architecture. Figure 4 presents the information processing diagram of the serial implementation of HICAT. In Figure 4 (as well as Figure 5), the vertical dimension roughly corresponds to time, flowing bottom up.

Stimulus S_i is first mapped onto an internal representation $\vec{\psi}_i$ by the box labeled *auditory processing*. Next, categorization A is completed in three steps. First, $\vec{\psi}_i$ is matched to stored category information, represented by category goodness functions $\gamma_{A_1}(\vec{\psi})$ and $\gamma_{A_2}(\vec{\psi})$ for categories A_1 and A_2, respectively. $\gamma_{A_1}(\vec{\psi})$ and $\gamma_{A_2}(\vec{\psi})$ are two-dimensional Gaussians on psychological space Ψ with equal covariance matrices C_A:

$$\gamma_{A_1}(\psi_x,\psi_y) = \exp -\frac{1}{2(1-\rho_A^2)}\left[\left(\frac{\psi_x-\mu_{A_1x}}{\sigma_{Ax}}\right)^2 - 2\rho_A\left(\frac{\psi_x-\mu_{A_1x}}{\sigma_{Ax}}\right)\left(\frac{\psi_y-\mu_{A_1y}}{\sigma_{Ay}}\right) + \left(\frac{\psi_y-\mu_{A_1y}}{\sigma_{Ay}}\right)^2\right], \quad (16)$$

where $\vec{\mu}_{A_1} = (\mu_{A_1x}, \mu_{A_1y})$ is the mean of γ_{A_1}, σ_{Ax} and σ_{Ay} are standard deviations of γ_{A_1} (and γ_{A_2}) along ψ_x and ψ_y, and ρ_A is the correlation coefficient of γ_{A_1} (and γ_{A_2}). An analogous expression holds for γ_{A_2}. Note that the "self-similarity" of the goodness functions equals one—that is, $\gamma_{A_1}(\mu_{A_1x}, \mu_{A_1y}) = 1$.

The matching process yields goodness values $\gamma_{A_1}(\vec{\psi}_i)$ and $\gamma_{A_2}(\vec{\psi}_i)$ for stimulus S_i. Next, the probabilities $p(A_1|S_i)$ and $p(A_2|S_i)$ of assigning labels L_{A_1} or L_{A_2} to stimulus S_i are calculated from the goodness values, using Luce's choice rule (Luce, 1963):

$$p(A_1|S_i) = \frac{\gamma_{A_1}}{\gamma_{A_1}+\gamma_{A_2}}, \quad (17)$$

and $p(A_2|S_i) = 1 - p(A_1|S_i)$. Finally, a choice is made between response labels L_{A_1} and L_{A_2} on the basis of $p(A_2|S_i)$ and $p(A_1|S_i)$, using the binomial distribution.

The secondary categorization B will start only when a discrete label is selected for the primary categorization. The details of the goodness match for categorization B depend on the outcome of categorization A. That is, the goodness functions for B_1 and B_2 are different for A_1 and A_2. The goodness functions for B_1 and B_2 given A_1 are indicated as $\gamma_{B_1|A_1}$ and $\gamma_{B_2|A_1}$, and analogously for A_2. The conditional probabilities for B are calculated by applying Luce's choice rule to the conditional goodness functions:

$$p(B_1|S_i,A_1) = \frac{\gamma_{B_1|A_1}}{\gamma_{B_1|A_1}+\gamma_{B_2|A_1}}, \quad (18)$$

with $p(B_2|S_i, A_1) = 1 - p(B_1|S_i, A_1)$, and analogously for A_2. Finally, the two labels are combined, and the output syllable label L_{AB} is generated.

The serial architecture is unattractive for two reasons. First, the output of the architecture will always be a discrete label. Although this is not a problem in the context of a phonetic categorization experiment (indeed, it is exactly what the task requires), it will raise problems in the context of everyday speech recognition. The reason for

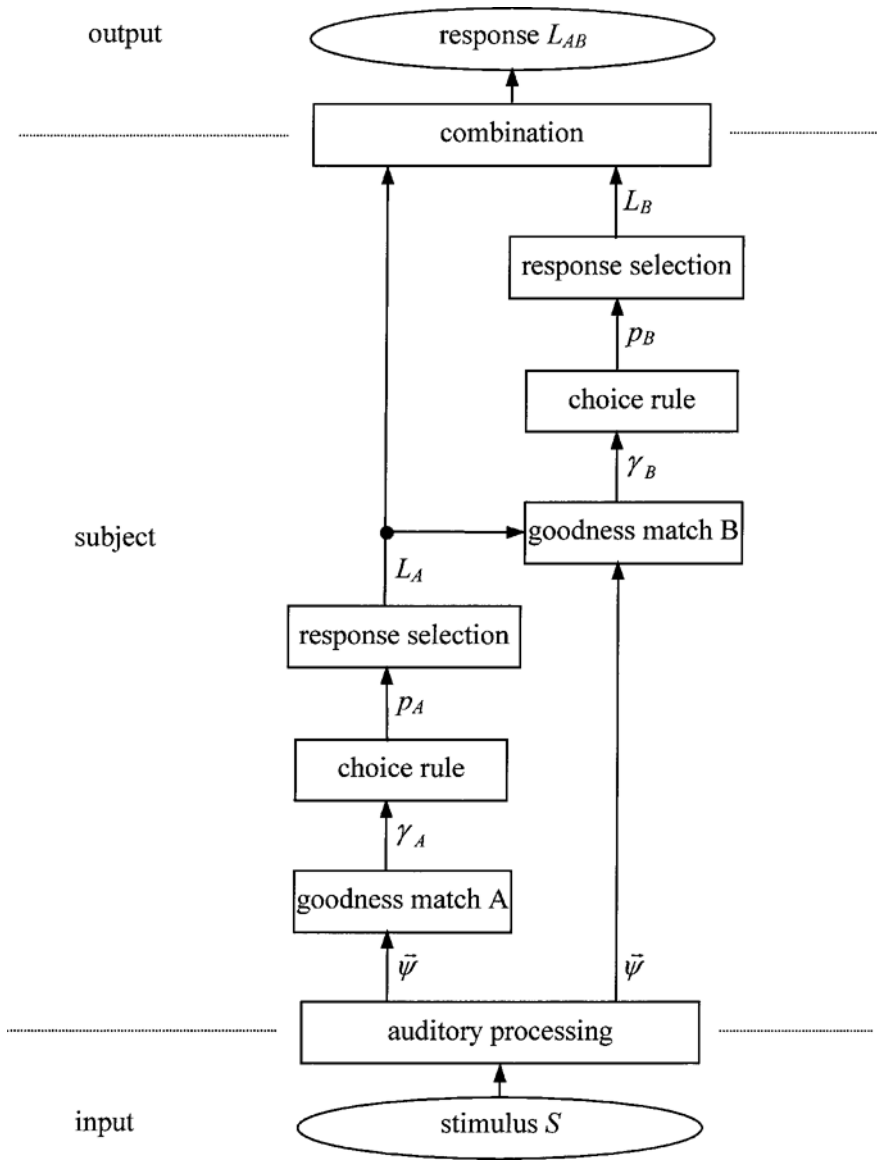

Figure 4. Information-processing diagram for the serial architecture of HICAT. The left-hand branch corresponds to categorization A, the right-hand branch to B. Dotted lines represent boundaries between the subject and the physical world. Ellipses represent directly observable information, rectangles represent processing modules, and arrows represent information flow. For the meaning of symbols, refer to the text.

this is that it is better if a classification process that involves multiple levels of representation—for example, phonemes and words—does not make hard, irreversible decisions at low levels prior to making such decisions at higher levels. It is better to retain phoneme probabilities or goodness values, so that all possible recognition results remain possible and the best one can be selected at a word or sentence level recognition (see Marslen-Wilson, 1987, and Nearey, in press, for more elaborate discussions).

The second reason the serial architecture is unlikely is the assumption that categorization B has to wait for A to finish before it can start. Several studies have suggested that this is not the way in which the human recognition system operates. For example, Repp (1980) and Whalen (1984) have presented evidence that phonetic information is continuously extracted from the speech signal and is mapped onto phonetic categories as soon as it becomes available. The parallel processing architecture presented below does not suffer from either of these drawbacks.

Parallel architecture. Figure 5 presents the information processing diagram of the parallel implementation of HICAT.

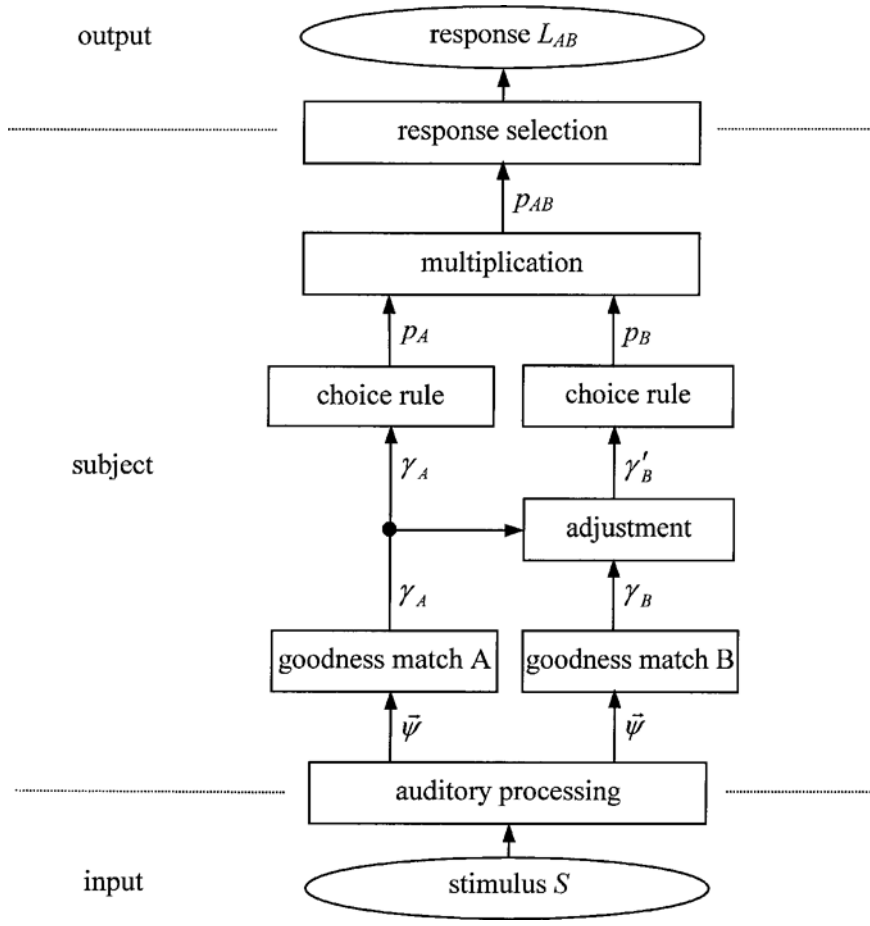

Figure 5. Information-processing diagram for the parallel architecture of HICAT. For further information, see Figure 4.

Essentially, both categorizations run simultaneously in the parallel architecture. As in the serial architecture, stimulus S_i is first mapped onto an internal representation $\vec{\psi}_i$. Next, $\vec{\psi}_i$ is matched to category representations $\gamma_{A_1}(\vec{\psi})$ and $\gamma_{A_2}(\vec{\psi})$ for A_1 and A_2, as well as to $\gamma_{B_1}(\vec{\psi})$ and $\gamma_{B_2}(\vec{\psi})$ for B_1 and B_2. γ_{A_1}, γ_{A_2}, γ_{B_1}, and γ_{B_2}, are the optimal goodness functions under the assumption of independence, with the additional assumption that the covariance matrices of A_1 and A_2 are equal and those of B_1 and B_2 are equal as well. Next, the goodness values $\gamma_{B_1}(\vec{\psi}_i)$ and $\gamma_{B_2}(\vec{\psi}_i)$ are adjusted to increase performance. The adjustment leads to four conditional goodness values for categorization B: $\gamma_{B_1|A_1}$, $\gamma_{B_2|A_1}$, $\gamma_{B_1|A_2}$, and $\gamma_{B_2|A_2}$. These values are indicated in Figure 5 as γ'_B. In the next step, Luce's choice rule is applied to γ_{A_1} and γ_{A_2} (see Equation 17) and to $\gamma_{B_1|A_1}$, $\gamma_{B_2|A_1}$, $\gamma_{B_1|A_2}$, and $\gamma_{B_2|A_2}$ (see Equation 18). Finally, the appropriate (conditional) probabilities are multiplied, yielding syllable probabilities, and a single syllable label L_{AB} is selected using the multinomial function.

It is assumed that during normal speech recognition, the goodness values γ_{A_1}, γ_{A_2}, $\gamma_{B_1|A_1}$, and so on, are passed on to higher level processes for recognizing words. Application of the choice rule plus subsequent response selection is thus assumed to be invoked by the experimental task.

The processing step that introduces the hierarchical dependency is the *adjustment* of the goodness values γ_{B_1} and γ_{B_2}. It is important to realize that the adjustment is not a form of feedback. There is simply a *lateral* information flow at the level of phoneme goodness values, so the architecture remains essentially bottom-up. It is also noted here that a parallel architecture with lateral information flow, such as the one presented in Figure 5, is capable of producing contextual effects of phonotactics (e.g., Massaro & Cohen, 1983), even though hard phoneme decisions are deferred as long as possible. The reason is that, as was explained above, all conditional phoneme probabilities remain available throughout and are combined appropriately at the decision stage, where one of the combinations is chosen. Consequently, the resulting probability functions for the parallel implementation are mathematically equivalent to those for the serial implementation.

Conceptually, the goodness value adjustments involve a shift of the goodness functions $\gamma_{B_1}(\vec{\psi})$ and $\gamma_{B_2}(\vec{\psi})$ in

psychological space, plus, in some cases, a bias. The shift may be along the axis joining the means of γ_{A_1} and γ_{A_2}, in which case the original (independent) functions γ_{B_1} and γ_{B_2} are multiplied or divided by a factor

$$\left(\frac{\gamma_{A_1}}{\gamma_{A_2}}\right)^q.$$

γ_{B_1} and γ_{B_2} may also be shifted along the axis joining the means of γ_{B_1} and γ_{B_2} themselves, in which case the factor is of the form

$$\left(\frac{\gamma_{B_1}}{\gamma_{B_2}}\right)^p.$$

In the case of *position dependency*, the mathematical definitions of the adjustments are the following:

$$\gamma_{B_1|A_1} = \gamma_{B_1}\left(\frac{\gamma_{B_1}}{\gamma_{B_2}}\right)^p\left(\frac{\gamma_{A_1}}{\gamma_{A_2}}\right)^q \exp(r) \tag{19}$$

$$\gamma_{B_2|A_1} = \gamma_{B_2}\left(\frac{\gamma_{B_1}}{\gamma_{B_2}}\right)^{-p}\left(\frac{\gamma_{A_1}}{\gamma_{A_2}}\right)^{-q} \exp(-r) \tag{20}$$

$$\gamma_{B_1|A_2} = \gamma_{B_1}\left(\frac{\gamma_{B_1}}{\gamma_{B_2}}\right)^p\left(\frac{\gamma_{A_1}}{\gamma_{A_2}}\right)^q \exp(-r) \tag{21}$$

$$\gamma_{B_2|A_2} = \gamma_{B_2}\left(\frac{\gamma_{B_1}}{\gamma_{B_2}}\right)^{-p}\left(\frac{\gamma_{A_1}}{\gamma_{A_2}}\right)^{-q} \exp(r). \tag{22}$$

The equations for *orientation dependency* are:

$$\gamma_{B_1|A_1} = \gamma_{B_1}\left(\frac{\gamma_{A_1}}{\gamma_{A_2}}\right)^q \exp(r) \tag{23}$$

$$\gamma_{B_2|A_1} = \gamma_{B_2}\left(\frac{\gamma_{A_1}}{\gamma_{A_2}}\right)^{-q} \exp(-r) \tag{24}$$

$$\gamma_{B_1|A_2} = \gamma_{B_1}\left(\frac{\gamma_{A_1}}{\gamma_{A_2}}\right)^{-q} \exp(r) \tag{25}$$

$$\gamma_{B_2|A_2} = \gamma_{B_2}\left(\frac{\gamma_{A_1}}{\gamma_{A_2}}\right)^q \exp(-r). \tag{26}$$

In the case of *steepness dependency*, the adjustment is defined mathematically as follows:

$$\gamma_{B_1|A_1} = \gamma_{B_1}\left(\frac{\gamma_{B_1}}{\gamma_{B_2}}\right)^p \tag{27}$$

$$\gamma_{B_2|A_1} = \gamma_{B_2}\left(\frac{\gamma_{B_1}}{\gamma_{B_2}}\right)^{-p} \tag{28}$$

$$\gamma_{B_1|A_2} = \gamma_{B_1}\left(\frac{\gamma_{B_1}}{\gamma_{B_2}}\right)^{-p} \tag{29}$$

$$\gamma_{B_2|A_2} = \gamma_{B_2}\left(\frac{\gamma_{B_1}}{\gamma_{B_2}}\right)^p. \tag{30}$$

Appendix B shows that Equations 19–30 lead to the actual HICAT model as defined in Equations 3–7. In addition, Appendix B lists the relations between the parameters p, q, and r to the pdfs of acoustical cues of Figures 1 and 2.

COMPARISON OF HICAT TO OTHER CATEGORIZATION MODELS

HICAT models processing dependencies in two binary categorizations. Three other models (or model classes) explicitly address this issue.

1. The fuzzy-logical model of perception (FLMP; Oden & Massaro, 1978), which is a special case of the recently proposed multinomial processing tree model (MPTM; Batchelder & Crowther, 1997).

2. The diphone-biased secondary-cue model (DBSCM; Nearey, 1990).

3. General recognition theory (GRT; Ashby & Townsend, 1986), a constrained version of which has been applied to phonetic perception by Kingston and Macmillan (1995).

The Fuzzy-Logical Model of Perception and the Multinomial Processing Tree Model

FLMP has been used to model 4RC data as well as two-response categorization (2RC) data (see Massaro, 1998). The present discussion is focused on the 4RC version of FLMP. The application of FLMP to the categorization of successive phonemes is limited (Massaro & Cohen, 1983). Still, it is very useful to compare the two models. FLMP distinguishes three processing steps. In the *evaluation* phase, relevant psychophysical features are extracted from the stimulus in a deterministic (i.e., noise-free) fashion. Each feature is matched to prototypical feature values for each category, resulting in a measure of category goodness, or *fuzzy truth*, for each feature for each of the categories. In the *integration* phase, the fuzzy-truth values for multiple features are combined through a multiplicative rule (or logical conjoining), resulting in a total category goodness of the stimulus for each of the relevant categories. Finally, in the *decision* phase, a single response is selected on the basis of the relative goodness of the stimulus for all categories, using Luce's choice rule.

The HICAT model is similar to FLMP in using essentially the same processing steps of deterministic stimulus encoding, followed by an integration phase resulting in a goodness value and, finally, by a stochastic response

selection. HICAT differs from FLMP on a number of points, however. The most fundamental difference concerns the assumptions underlying the integration of information from separate stimulus dimensions. This is illustrated in Figure 6.

In FLMP (left-hand panel of Figure 6), the physical stimulus dimension ϕ_x is mapped directly onto goodness levels γ_A and $1 - \gamma_A$ for categories A_1 and A_2, and ϕ_y is mapped onto goodness levels γ_B and $1 - \gamma_B$ for categories B_1 and B_2. Next, the appropriate goodness levels are multiplied, yielding syllable goodness levels $\gamma_A\gamma_B$, $(1 - \gamma_A)\gamma_B$, and so on, and Luce's choice rule is applied to the syllable goodness levels. These assumptions imply that the four syllable prototypes form the corners of a rectangle in psychological space, with sides parallel to the stimulus dimensions. Thus, in FLMP each stimulus dimension is related to one phoneme only. In HICAT (right-hand panel of Figure 6), on the other hand, physical stimulus dimensions ϕ_x and ϕ_y are initially mapped onto psychological dimensions ψ_x and ψ_y. Next, goodness levels for categories A_1 and A_2 are calculated on the basis of *both* psychological dimensions. The same holds for B_1 and B_2. Next, the goodness levels are converted into phoneme probabilities separately for A and B, after which the probabilities are combined, through multiplication, to form syllable probabilities. So Luce's choice rule is applied at the phoneme level in HICAT, not at the syllable level.

These differences in processing assumptions in FLMP and HICAT have important consequences. First of all, in its usual 4RC implementation, FLMP is restricted to having category boundaries that are parallel to the physical stimulus dimensions (as also was observed by Batchelder & Crowther, 1997). In other words, FLMP cannot model situations with significant sharing of acoustic cues. This restriction does not apply to HICAT. It is confusing, however, that this restriction also does *not* apply to the two-alternative forced-choice implementation of FLMP. It remains, therefore, unclear whether the assumption, implicitly made in the 4RC implementation, that cue sharing is not allowed is essential to the FLMP framework.

Early versions of FLMP (Massaro & Cohen, 1983; Massaro & Oden, 1980; Oden & Massaro, 1978), henceforth indicated as *FLMPmod*, incorporated so-called *feature modifiers* in the prototype definitions. The modifiers were introduced to allow for the possibility that features may have more extreme values for some prototypes than for others. For example, voice onset time (VOT), which is a primary cue for stop voicing, is known to play a secondary role in stop place recognition, where /d/ is expected to have a longer VOT than /b/. Mathe-

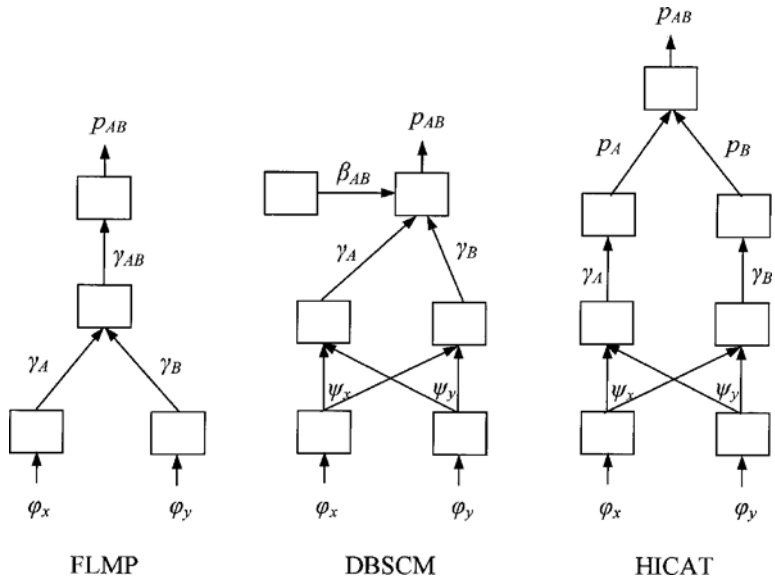

Figure 6. Comparison of information processing diagrams for the four-response categorization fuzzy-logic model of perception (FLMP, left), the diphone-biased secondary-cue model (DBSCM, middle), and HICAT (right). Rectangles represent processing modules; arrows represent information flow. Note that the goodness functions depend on one physical variable only in FLMP versus a dependence on both in DBSCM and HICAT, and note how information for two phonemes is combined at the level of the goodness function in FLMP and DBSCM versus at the probability level in HICAT.

matically, the modifiers are implemented by using exponentials in the prototype definitions. The fuzzy match $\gamma(S_{ij})$ of stimulus S_{ij} to the four prototypes A_1B_1, A_1B_2, A_2B_1, and A_2B_2 can be expressed as follows:

$$\gamma_{A_1B_1}(S_{ij}) = a_i^{m_1} b_j^{m_2} \tag{31}$$

$$\gamma_{A_1B_2}(S_{ij}) = a_i^{m_3} \left(1-b_j\right)^{m_4} \tag{32}$$

$$\gamma_{A_2B_1}(S_{ij}) = \left(1-a_i\right)^{m_5} b_j^{m_6} \tag{33}$$

$$\gamma_{A_2B_2}(S_{ij}) = \left(1-a_i\right)^{m_7} \left(1-b_j\right)^{m_8}, \tag{34}$$

where a_i and b_j indicate psychological effects of two independently varied stimulus parameters and m_k represents feature modifier k. As in the basic FLMP formulation, Luce's choice rule is applied to the goodness levels $\gamma(S_{ij})$. When one or more feature modifiers are different from unity, the categorizations of A and B are no longer independent, because now

$$\begin{aligned} p\left(A_1 \mid S_{ij}\right) p\left(B_1 \mid S_{ij}\right) = & \left[p\left(A_1B_1 \mid S_{ij}\right)\right. \\ & \left. + p\left(A_1B_2 \mid S_{ij}\right)\right]\left[p\left(A_1B_1 \mid S_{ij}\right)\right. \\ & \left. + p\left(A_2B_1 \mid S_{ij}\right)\right] \end{aligned} \tag{35}$$

$$\neq p\left(A_1B_1 \mid S_{ij}\right). \tag{36}$$

When a modifier is applied to a prototype definition, the location of the prototype is unchanged—that is, the four prototypes still form a rectangle. The modifier only affects the *spread* of the goodness function around the prototype. Modifier values larger than one make the goodness function narrower, thus requiring feature values closer to the prototype for perceiving the associated category.

Although, thus described, the effect of the modifiers is transparent and appealing, it is difficult to gauge the added modeling capacity of FLMPmod in terms of constraints on categorization function positions, orientations, and steepnesses, especially when the full set of eight modifiers is used, as in Oden and Massaro (1978) and Massaro and Oden (1980). This makes a comparison of FLMPmod with HICAT difficult. However, given certain constraints on the feature modifiers, FLMPmod can model certain hierarchical dependencies in a similar, although not identical, fashion to HICAT. Dependencies of the position and steepness of the B_1–B_2 categorization function on categorization A can be modeled with three free parameters, as follows:

$$\gamma_{A_1B_1}(S_{ij}) = a_i b_j^{m_0+m_1+m_2} \tag{37}$$

$$\gamma_{A_1B_2}(S_{ij}) = a_i \left(1-b_j\right)^{m_0-m_1+m_2} \tag{38}$$

$$\gamma_{A_2B_1}(S_{ij}) = \left(1-a_i\right) b_j^{m_0-m_1-m_2} \tag{39}$$

$$\gamma_{A_2B_2}(S_{ij}) = \left(1-a_i\right)\left(1-b_j\right)^{m_0+m_1-m_2}, \tag{40}$$

where m_0 is the average modifier value and m_1 and m_2 are associated with *position* and *steepness* dependency, respectively. Note that the modifiers apply only to categorization B. With respect to *orientation* dependence, it is noted that Oden and Massaro (1978, Figure 5, p. 185) illustrated that the general definition of FLMPmod with eight modifiers (Equations 31–34) can model category boundaries that are neither parallel to the stimulus dimensions nor necessarily mutually parallel (as in Nearey's DBSCM). I have, however, not been able to find constrained expressions for FLMPmod in the vein of Equations 37–40 in which a single parameter captures a hierarchical boundary orientation dependency. As we saw earlier for the HICAT model, the present analysis of FLMP with feature modifiers shows once again that it is possible to have (hierarchical) categorization dependencies without violating the autonomy of the categorization process or the independence of the initial fuzzy phoneme or feature categorization.

Besides differing in assumptions about information processing, FLMP and HICAT support different recognition units. Massaro supports the syllable as the basic recognition unit (e.g., Massaro, 1998). Indeed, in FLMP's applications to categorization of successive phonemes, the category prototypes are syllable prototypes (Massaro & Cohen, 1983). However, FLMP's integration rule via a simple multiplication of phoneme goodness levels is essentially equivalent to an assumption of independence of phoneme decisions, which suggests that, actually, the phoneme would be a good candidate for a recognition unit (see Nearey, in press, for a similar argument). In HICAT, the unit of recognition is defined as the unit at which the goodness functions are defined (and stored in memory), which is the phoneme.

A related difference between the two models concerns the information contained in the actual category representation. In FLMP, this is a syllable prototype, of the form "low ϕ_x and high ϕ_y," and so on, where ϕ_x and ϕ_y are physical stimulus parameters. In HICAT, on the other hand, the category representation is a multidimensional Gaussian goodness function. Besides the *location* of a prototype (i.e., the mean of the distribution), such a function also contains information about its *spread* in the form of variances and covariances around the mean.

In HICAT, it is assumed that the stored goodness functions representing the categories are equal to the (unnormalized) pdfs of acoustic cues in natural speech. The relation of the model parameters of individual FLMP fits to cue distributions in natural speech has not featured prominently in the work by Massaro and colleagues. In a qualitative sense, it is claimed that, through exposure to speech, infants quickly acquire the prototypes necessary for speech recognition and that infants raised in different language environments have different prototypes (Massaro, 1998). Oden (1992) proposed a learning procedure using error backpropagation, which enables FLMP to learn its fuzzy propositions from individual training trials.

Finally, FLMP and HICAT differ in the way the mapping of the physical stimulus parameters to the goodness

functions is usually implemented. In FLMP's standard implementation, separate parameters are assigned to each level of each stimulus dimension. This makes FLMP mathematically equivalent to a log-linear model (Crowther, Batchelder, & Hu, 1995). In contrast, HICAT assumes that the goodness functions are Gaussian, which leads to continuous logistic categorization functions. Inspection of FLMP parameters reported in studies by Massaro and colleagues shows that they always exhibit sigmoid-like shapes, which gives some support for the constraint made in HICAT. It should be mentioned, however, that Oden and Massaro (1978) did use a continuous physical-to-psychological mapping function in one of their model analyses. Application of this constraint worsened the goodness of fit only by a very small amount. So, although the constrained stimulus encoding in FLMP has, to my knowledge, never been used since, it can be argued that the usual choice of discrete stimulus encoding is not so much a core processing assumption of FLMP as a mere implementation issue. Nevertheless, with respect to the standard implementation of FLMP, it is useful to note that, in one sense, HICAT is more constrained than FLMP by assuming logistic categorization functions where FLMP allows for arbitrary categorization functions. At the same time, FLMP is more constrained than HICAT, in the sense that it assumes independence, both acoustically and phonologically, where HICAT allows for cue sharing and hierarchical dependency. HICAT is most similar to 4RC FLMP when $p_y = q_x = 0$ (no acoustic context) and $c_0 = c_\alpha = c_\beta = 0$ (no phonological context), leaving only four free parameters.

Recently, Batchelder and Crowther (1997) introduced the family of MPTMs, of which FLMP is a special case. In MPTM, FLMP is extended in two ways: (1) Both categorizations are allowed to depend on both stimulus dimensions, and (2) one of the categorizations is allowed to depend on the outcome of the other. Essentially, these two types of processing dependencies are exactly those included in HICAT. HICAT is, therefore, mathematically very similar to MPTM, except that HICAT assumes Gaussian goodness functions, resulting in logistic categorization functions, whereas MPTM, like FLMP, assumes nominal stimulus levels. In practice, this difference is important, because HICAT's constraint keeps the number of parameters down to a maximum of nine, whereas especially Generalization 1, given above, leads to a proliferation of parameters to twice the number of parameters in FLMP or more.

An additional important difference between MPTM and HICAT is that HICAT is proposed as a genuine psychological model of categorization. It not only includes explicit descriptions of hypotheses regarding the various processing stages in categorization, it also can be used to predict processing dependencies on the basis of acoustic measurements on natural utterances (as will be more fully discussed below). In contrast, Batchelder and Crowther's presentation of MPTM is more statistically than psychologically oriented, and the aim of the model is analysis only. MPTM is a very useful generalization of FLMP, however, because it allows for explicit statistical testing of various processing hypotheses that FLMP does not incorporate.

Diphone-Biased Secondary-Cue Model

For more than a decade, Nearey (1990, 1997, in press) has supported a model of speech perception in which the phoneme is the unit of recognition and the classifications of successive phonemes are phonologically independent. The model deviates from the strongest form of independence, as represented by, for example, FLMP, in two ways: Cue sharing (using *secondary cues* in Nearey's terminology) is allowed, as are so-called diphone biases. Diphone biases favor particular combinations of phonemes over others; for example, given a choice between /si/, /su/, /ʃi/, and /ʃu/, /su/ and /ʃi/ are generally more likely than /si/ and /ʃu/. Such preferences are genuine response biases because they do not depend on stimulus information. Because of these core assumptions, Nearey's model is referred to as the DBSCM.

The middle panel of Figure 6 illustrates the processing architecture of the 4RC version of DBSCM. As in HICAT, the first processing step involves a mapping of physical stimulus parameters ϕ_x and ϕ_y to psychological parameters ψ_x and ψ_y. Next, phoneme goodness values are calculated from the two stimulus dimensions. Finally, diphone probabilities are calculated on the basis of the phoneme goodness values and the relevant diphone biases, indicated in Figure 6 as β_{AB}.

As Figure 6 illustrates, the initial stages of the physical-to-psychological mapping and goodness calculation are identical in HICAT and DBSCM. They diverge in the subsequent processing of the goodness values. Whereas in HICAT there is a one-way lateral influence among the goodness values, they remain independent in the DBSCM. Furthermore, in HICAT individual phoneme response probabilities are calculated, whereas in DBSCM the phoneme goodness values, together with the diphone biases, are transformed to syllable probabilities in one step. However, if diphone biases or higher order interactions are excluded from the DBSCM, it is mathematically equivalent to HICAT without hierarchical dependencies ($c_0 = c_\alpha = c_\beta = 0$).

The effect of HICAT's position dependency, coded by parameter c_0, is similar to that of the diphone bias in DBSCM. In a territorial plot, both have the effect of creating a section of the space where two diagonally opposed categories (e.g., *bat* and *bed* in Experiment 1 of Whalen, 1989) share a border, whereas the other two categories never touch. Compare, for example, Figure 3 in Nearey (1990) to Figure 3B in the present paper or Figure 2 in Smits (in press). HICAT's other two dependency parameters (c_α and c_β) have no analogues in DBSCM.

The psychological reality and function of the diphone bias have been subjected to discussion. Whalen (1992) criticized the diphone bias for making the DBSCM too unconstrained. No predictions are made concerning which diphones will actually be favored over which oth-

ers. Kluender and Lotto (1999), on the other hand, argued that the DBSCM with diphone bias is sensibly constrained in the sense that it supports only territorial map structures in which only one response region is associated with each response—that is, XOR-type classifications cannot be modeled. Nearey (in press) himself has offered the following potential interpretations of the diphone bias. First, it may code likelihoods of co-occurrence of diphones and phonotactic constraints. In a given language, some phoneme combinations are more common than others, which can be effectively captured by diphone biases. Second, they may be employed to approach optimal recognition, given natural cue distributions. It can be shown that the optimal unconditional classification of the shifted geometry (Figure 1B) involves the use of a diphone bias.

Nearey (1990, 1997) has used logistic regression (LR) as a statistical framework to test for the hypotheses incorporated in DBSCM. In LR, the influence of various experimental parameters on frequencies in a stimulus–response matrix are systematically decomposed into main stimulus effects, main response effects, and various interactions, in a fashion similar to the analysis of variance model, allowing for statistical testing of the significance of each of the terms. The *saturated* version of the LR model includes terms that are equivalent to phonological dependencies. If such *stimulus-tuned diphone* terms would prove to be significant in an LR analysis of a set of categorization data, this would be interpreted by Nearey as evidence for a recognition unit larger than the phoneme. In contrast, HICAT allows for three types of phonological context, while the phoneme unit is still adhered to. Importantly, however, the dependencies in HICAT, being hierarchical, are different from the stimulus-tuned diphone terms, which are symmetrical—that is, nonhierarchical—in nature.

Nearey has always emphasized the importance of scaling up the issues of the recognition unit and processing dependencies from individual model fits to a realistic speech recognition situation involving all phonemes in a particular language (see, especially, Nearey, in press). Nearey's argument for allowing diphone biases but disallowing stimulus-tuned diphone terms was centered around a reflection on the complexity of the speech recognition system, as quantified by its number of parameters. In Nearey's approach, every diphone bias leads to one extra free parameter in the speech recognition system, which would add, at most, a few thousand parameters to an independent phoneme recognizer. Inclusion of stimulus-tuned diphone terms, on the other hand, would add a number of parameters equal to the number of diphones (a few thousand) times the number of cues for each diphone (perhaps, on the order of 10?). It is important to note that hierarchical dependencies, as they are implemented in HICAT, do not lead to such an undesirable increase in processing complexity. Like Nearey's diphone bias, each dependency parameter would, at most, lead to a few thousand extra parameters in the full system.

Finally, it is worth noting that in its first formulation (Nearey, 1990), the DBSCM was related to the NAPP model of Nearey and Hogan (1986). Nearey (1990) indicated that the 4RC version of DBSCM is equivalent to an optimal fuzzy classifier if the natural cue distributions are multidimensional Gaussian with equal covariance matrices for the four syllables. As was mentioned earlier, Nearey (1992) qualitatively discussed how the variously constrained versions of the model would deal with the classification of the syllables /si/, /su/, /ʃi/, and /ʃu/, given the natural cue values derived from a study by Soli (1981). Nearey never attempted, however, to quantitatively link the observed categorization patterns to acoustical measurements on naturally produced speech. Nevertheless, it is important to note that the diphone term gets a new interpretation, not previously proposed by Nearey, when considered in the context of such cue geometries. In fact, it can be shown that Nearey's diphone-biased model gives an optimal (unconditional) fuzzy categorization of the *shifted* geometry of Figure 1B. (Recall that HICAT gives the optimal *conditional* categorization.) Nearey has proposed an interpretation of the diphone bias as capturing phoneme transition probabilities (Nearey, 1990, 1997). Interestingly, the same set of parameters would therefore code shifts in acoustical cue distributions owing to coarticulation and phoneme transition probabilities.

General Recognition Theory

GRT (Ashby & Townsend, 1986) is a multidimensional signal-detection model. The perceptual effect of repeated presentations of a stimulus is modeled by a multidimensional (usually Gaussian) pdf on a psychological space. The space is divided into regions, each of which corresponds to a response. On presentation of a stimulus, the stimulus is mapped onto a point in perceptual space, in which region the point is located is evaluated, and the associated response is emitted. Because GRT is a relatively unconstrained framework for categorization theories incorporating a variety of potential processing dependencies, an information processing architecture of GRT in the context of speech perception (in the vein of Figure 6) is not available.

Essential to the GRT framework is the distinction between (in)dependence at the perceptual versus decisional stages, where the perceptual stage refers to stimulus encoding and the decisional stage to response selection. Originally, Ashby and Townsend (1986) distinguished five types of independence in categorization within the GRT framework. Two of those, perceptual and decisional separability, have been featured in subsequent work by Ashby and Maddox (e.g., Maddox, 1992) and have been the focus in the application of GRT to speech perception (e.g., Kingston & MacMillan, 1995). I will therefore limit the discussion to perceptual and decisional separability.

Perceptual separability holds if the perceptual effect of one variable does not depend on the value of another variable. In an orthogonal combination of two binary variables A and B, the marginal distribution of perceptual

effects of stimulus A_1B_1 along dimension A should be identical to that of stimulus A_1B_2, and so forth. In terms of the fricative–vowel example used earlier, perceptual separability holds if a rectangular stimulus array in the physical $F3 \times F_{fr}$ space is mapped onto a rectangular array in psychological space. If, for example, a rectangular stimulus array in physical space is instead mapped onto a parallelogram in psychological space, perceptual separability is violated. This particular type of violation is indicated as *mean-shift integrality* (Maddox, 1992). Kingston, Macmillan, and colleagues (Kingston & Macmillan, 1995; Kingston, Macmillan, Walsh-Dickey, Thorburn, & Bartels, 1997; Macmillan, Kingston, Thorburn, Walsh-Dickey, & Bartels, 1999) carried out detection-theory analyses of vowel categorization data produced with the Garner paradigm. The studies showed that, in listeners' categorizations of English vowels, mean-shift integrality applies to acoustic dimensions associated with separate articulations.

Decisional separability refers to the response selection mechanism only. Decisional separability holds if the decision about one component of the stimulus is independent of the value of the other. This is equivalent to the decision bounds' being parallel to the coordinate axes. In terms of the fricative–vowel perception, decisional separability holds if the perceived fricative is unaffected by $F3$ and the vowel is unaffected by F_{fr}, as in Figure 2A.

An important difference between HICAT and GRT concerns the assumptions about the locus of the stochastic component. This makes it difficult to find equivalents for every conceivable type of independence in both models. However, it is relatively straightforward to translate perceptual and decisional separability to the HICAT framework. HICAT allows for violations of both perceptual and decisional separability. As was discussed earlier, one of the assumptions underlying HICAT is that cue sharing is common in speech perception and, accordingly, decision bounds are allowed to have any orientation. HICAT can indeed be used to test for decisional separability by testing for the significance of parameters p_x, p_y, q_x, and q_y in Equations 3 and 4. HICAT cannot, however, be used to test for perceptual separability. As in GRT, violation of perceptual separability in HICAT is equivalent to the mapping of a rectangular stimulus array onto a nonrectangular one. However, we need to distinguish between distributions of the physical parameter values found in natural speech, as represented in Figures 1 and 2, and distributions of the perceptual effects of repeatedly presented stimuli. If in HICAT decision bounds are found to be nonparallel to the coordinate axes, this may be due either to violations of perceptual separability or to the classifier's having chosen the boundary orientations to deal with natural cue distributions that are themselves not arranged in a neat rectangle. Of course, the two may even be superimposed.

Distinguishing Competing Models Empirically

How can HICAT, FLMP, DBSCM, and GRT be distinguished on the basis of experimental data? Massaro and Friedman (1990) and Cohen and Massaro (1992) compared several models of information integration to FLMP, including GRT, multidimensional scaling (MDS; e.g., Shepard, 1962), and certain connectionist models. They compared 2RC and 4RC versions of the models theoretically, as well as empirically, by evaluating model fits to simulated FLMP-generated data and to experimental data. Both studies showed that, despite important differences in the assumptions on the locus of the stochastic component and on the decision mechanism, many of the models could be reduced to a likelihood-product form and are, thus, mathematically equivalent and indistinguishable on the basis of standard factorial categorization data. Cohen and Massaro suggested that alternative dependent measures, such as ratings, similarity judgments, and reaction times, might allow the models to be discriminated empirically.

These results should be put in perspective, however. As was discussed earlier, category prototypes in the 4RC version of FLMP are always assumed to form the four corners of a rectangle in psychological space, with the sides parallel to the psychological axes corresponding with the physical stimulus dimensions. In Massaro and Friedman (1990) and Cohen and Massaro (1992), this constraint was imposed on the 4RC versions of all alternative models, although they are not in any way "inherent" to the models. Therefore, the general conclusion reached in both studies that most models are mathematically equivalent, although appropriate for the 2RC versions, is too strong for the 4RC case. The correct conclusion for the 4RC models would be that certain constrained variants of the models are mathematically equivalent to FLMP. In particular, in the MDS and SDT/GRT models, the prototype-rectangle constraint is neither in any way built in nor commonly made in practice. In their usual application, these models allow category prototypes to be located anywhere in psychological space.

An empirical comparison of the categorization models discussed above should focus on four aspects: (1) the locus of the stochastic component, (2) acoustical dependencies (cue sharing), (3) (hierarchical) phonological dependencies, and (4) the linking of category representation and processing mechanisms to distributions of natural data. Concerning the locus of the stochastic component, it is important to realize that although the theoretical assumptions of stochastic stimulus encoding versus stochastic choice are fundamentally different, they generally lead to similar or even identical predictions with respect to categorization functions (see also Massaro & Friedman, 1990). Thus, the two hypotheses are in practice difficult to discriminate on the basis of categorization data. If the extra assumption is made that the category goodness values evoked by a stimulus are equivalent to the probability densities of the natural cue distributions, testing becomes easier. However, reported results are still equivocal. Ashby and Gott (1988), for example, found evidence for the use of hard decision criteria in basic visual categorization, whereas Lee and Zen-

tall (1966) found support for Luce's choice rule. In the context of phonetic categorization, Nearey and Hogan (1986) found that both hypotheses gave good descriptions of listeners' categorizations of voicing in Thai stops. Furthermore, Nearey and Assman (1986) and Andruski and Nearey (1992) showed that Luce's choice rule, applied to natural distributions of acoustic dimensions of vowels, leads to accurate predictions of listeners categorizations of the vowels, but similar results may be obtained in a GRT framework. In any case, the issue of the locus of the stochastic component in phonetic categorization is far from settled.

It is proposed here that a possible test between the two alternatives may be based on so-called range effects—in particular, the influence of variation of the size of a stimulus continuum on categorization functions (e.g., Parducci, 1965). The trace-context model of SDT incorporates range effects and makes very explicit predictions about the influence of stimulus range on discrimination and identification performance (Durlach & Braida, 1969; Macmillan, Goldberg, & Braida, 1988). In contrast, a strict version of Luce's choice rule applied to stimulus distributions, as used in HICAT, predicts the absence of range effects in phonetic categorization. Range effects have indeed been reported for phonetic categorization, but they are often small (Repp & Liberman, 1987). Of course, SDT-type perceptual noise associated with context variance could be straightforwardly incorporated in a categorization model like HICAT. Such a hybrid model would predict a shallower categorization function with increasing continuum size, thus incorporating range effects. However, an empirical finding of a phonetic categorization function that is *steeper* than would be predicted from Luce's choice rule applied to natural cue distributions could not be handled by such a hybrid model, because adding a stochastic component to the stimulus encoding stage can only make categorization functions shallower. To my knowledge, such findings have not yet been reported.

The second potential diagnostic discriminating between the various models concerns the presence or absence of acoustic context effects. This issue can be directly addressed by using two-dimensional stimulus continua, where one acoustic dimension, which is presumably associated with one phoneme distinction, is orthogonally crossed with another dimension associated with an adjacent phoneme. It is now generally accepted that phonetic categories are essentially multidimensional and that human categorization of every conceivable phonetic categorization is influenced by many acoustic parameters (e.g., Diehl & Kluender, 1987; Lisker & Abramson, 1970). More important, there is ample evidence that cues are shared between categorizations of successive phonemes. Whalen (1989) showed how the categorization of both the vowel and the final consonant depended on the duration and frequency of the first formant in /bVC/ words. Nearey (1997) conducted an extended version of Whalen's experiment, orthogonally varying four acoustic parameters and employing 10 response categories. As for Whalen's experiment, the results left no doubt that cues were shared between vowel and consonant judgments. As was indicated earlier, 4RC FLMP without the feature modifiers cannot account for such acoustic context effects.

The third aspect on which the models can be distinguished concerns (hierarchical) phonological context effects. Essentially, HICAT allows for such dependencies, whereas the other models do not, although DBSCM and GRT can model diphone biases, which, although theoretically distinct, are expected to be experimentally difficult to distinguish from HICAT's shift dependency. The little data relevant to the issue of phonological dependencies is equivocal. Smits (in press) applied HICAT to the data sets from Experiments 1 and 3 of Whalen (1989). Although good DPSCM fits of these data sets had been reported by Nearey (1990), Smits's analyses provided an alternative interpretation involving hierarchical dependencies. Goodnesses of fit of the two models were comparable, however, so there was no basis for a choice between the models. Smits (in press) subsequently presented new data from an experiment in which listeners categorized stimuli from a two-dimensional stimulus continuum as /si/, /sy/, /ʃi/, or /ʃy/. HICAT model analyses showed clear evidence for a dependency of the /s/–/ʃ/ boundary on the perceived vowel. Although HICAT fitted the new data better than did DPSCM, a direct statistical comparison of the two models was hindered by the fact that they are not nested. In the future, cross-validation techniques may offer a general solution here. Alternatively, given the mathematical similarity of HICAT and DPSCM, it should not be too difficult to construct a "supermodel" that includes both models as a special case, thus allowing for systematic significance testing of various parameters. The supermodel approach is more difficult for comparisons of HICAT and FLMP, or HICAT and GRT, because the mathematical structures of the models are so different.

A fourth way of distinguishing the models may concentrate on how well they predict various aspects of the categorization of coarticulated phonemes on the basis of natural distributions of relevant acoustical cues. Currently, HICAT is unique in making such predictions both qualitatively and quantitatively. Again, future comparison with DPSCM should be easiest, because the prediction method presented in a later section can probably be adapted for this model.

Finally, as has been suggested by Massaro and Friedman (1990), reaction times may provide an alternative empirical basis for discriminating the models. At present, only FLMP and GRT explicitly model the time course of categorization (Ashby & Maddox, 1994; Massaro & Cohen, 1991). HICAT and DPSCM only model categorization probabilities and need to be fleshed out to incorporate reaction times. In particular, hypotheses need to be formulated about the potential influence of time pressure on the size of the dependencies in HICAT. Does the *lateral* influence of one phoneme categorization on the other take time to build up? Similarly, the applica-

tion of diphone biases in DPSCM may have a temporal component.

MODEL SELECTION PROCEDURE

One of the obvious uses of the HICAT model is to determine, in cases of hierarchical dependencies, what the dependency *direction* is. The simplest approach to this issue would be to fit HICAT on the data in both ways, one with categorization A depending on B and one the other way around, and then selecting the model that gives the better fit. This is generally not a reliable approach. Owing to the stochastic nature of the categorization data, the model representing the incorrect direction may accidentally give a better fit than the correct one, especially for small data sets. It is therefore important to develop some sense of the reliability of a specific model choice, given a data set. I decided to test the reliability of the model choice through Monte Carlo simulations. The following parameters were thought to be potentially of influence on the probability of making the correct choice and were therefore varied in the simulations: number of stimuli $N_s = 25\ (5 \times 5)$, $100\ (10 \times 10)$; number of presentations of each stimulus $N_p = 25, 100$; position dependency parameter $c_0 = 0, 0.2, 0.8$; orientation dependency parameter $c_\alpha = 0, 0.05, 0.2$; steepness dependency parameter $c_\beta = 0, 0.05, 0.2$; size of vectors and $\vec{p}$ and $\vec{q}$ = 3, 12; and angle between vectors $\vec{p}$ and $\vec{q}$ = 45°, 90°.

These seven parameters were varied orthogonally in the simulations. Leaving out the cases in which $c_0 = c_\alpha = c_\beta = 0$, this led to a total of 416 parameter settings. For each parameter setting, the HICAT model was used to generate 10 different sets of categorization data. On each of the 4,160 resulting data sets, the HICAT model was fitted for both dependency directions, and the G^2 values for the two dependency directions were recorded, with lower G^2 corresponding to better fit (e.g., Agresti, 1990). It was evaluated whether the correct model indeed gave the lower G^2 .

Next, LR was carried out, with a parameter indicating whether the correct dependency direction was found to be a dependent variable and various combinations of the seven parameters listed above, plus the absolute difference in G^2 for the two directions ($|\Delta G^2|$), to be independent parameters. The analyses showed that a good prediction of the probability p_c of choosing the correct direction could be made on the basis of $|\Delta G^2|$ alone. The prediction formula is

$$p_c = \frac{1}{1+\exp\left(-0.38*|\Delta G^2|\right)}. \qquad (41)$$

Equation 41 predicts, for example, that a difference in G^2 between the fits for the two dependency directions of 10 or 20 leads to probabilities of correct choice of p_c = .98 and p_c = .9995, respectively. Of course Equation 41 is based on the assumption that one of the models is in fact the true model. In practice, categorization data are a good deal noisier than those in the Monte Carlo simulation, and G^2 values can become rather large. Therefore, it was thought useful to repeat the process for the absolute difference in G^2 for the two directions, *relative to* either the largest value $G^2_{\max}$ of the two

$$\left(\frac{|\Delta G^2|}{G^2_{\max}}\right)$$

or the smallest

$$\left(\frac{|\Delta G^2|}{G^2_{\min}}\right).$$

The variable

$$\left(\frac{|\Delta G^2|}{G^2_{\max}}\right)$$

turned out to be the better predictor, especially when combined with the number of stimuli N_s. The resulting prediction formula is

$$p_c = \frac{1}{1+\exp\left(-0.59*N_s*\frac{|\Delta G^2|}{G^2_{\max}}\right)}. \qquad (42)$$

In short, the proposed procedure for fitting HICAT to a data set is as follows. First, the independent model is fitted, setting c_0, c_α, and c_β to zero. Next, the full HICAT model, including c_0, c_α, and c_β, is fitted for the two dependency directions. If one or both of the resulting G^2 values are sufficiently different from G^2 for the independent model, it is concluded that a dependency is present. Next, the best-fitting dependency direction is chosen, and Equation 42 can be used to estimate the probability that this choice is actually correct. (Equation 41 can be used on the rare occasions when G^2 values are low enough to conclude that there is no significant lack of fit.) Finally, statistical tests can be carried out on the selected model to establish whether parameters can be left out without significant loss of fit.

PREDICTING PROCESSING DEPENDENCIES FROM ACOUSTICAL CUE DISTRIBUTIONS

As was discussed earlier, one of the basic claims of the pattern classification approach to speech perception is that important aspects of listeners' categorization of speech sounds are tuned to the statistical properties of the acoustic material that serves as input to the system. Starting from this claim, the HICAT model was defined, and the previous sections have focused on the mathematical structure of the HICAT model, why it represents a useful categorization strategy, and how it can be used to infer processing strategies from a set of categorization data. The present section presents an even stronger test of the above claim. A method is proposed for *predicting* likely categorization dependencies, given a set of acoustic measurements on natural utterances.

In natural data, we will generally not encounter pdf geometries as neat as the basic ones presented in Figure 1. Even when the assumed relations between covariance matrices more or less hold, the obtained geometries may be rotated, stretched, and/or mirrored versions of those in Figure 1. The method presented below starts with two successive linear transforms that remove such linear distortions from the data and make the problem identical to that of Figure 1 (after application of the same linear transforms). Next, hierarchical processing assumptions are straightforwardly translated into assumptions on the means and covariance matrices of phonemes B_1 and B_2. These assumptions can, in turn, be translated into conditions on "inferred" syllable distributions. Given these conditions, the optimal conditional categorization model can be derived.

Transform 1

Transform 1 "whitens" the pdfs of the phonemes $A_1$and A_2, so that their covariance matrices become identity matrices (e.g., Fukunaga, 1990).

First, means $\vec{\mu}_{A_1}$ and $\vec{\mu}_{A_2}$ and covariance matrix C_A for categories A_1 and A_2 are estimated from the acoustical data (after appropriate transforms to psychological axes). Let the N_t training vectors for the syllable A_1B_1 in psychological space Ψ be indicated by $\vec{\psi}^{\,i}_{A_1B_1}$, $i = 1..N_t$ and similarly for the other syllables. $\vec{\mu}_{A_1}$ and $\vec{\mu}_{A_2}$ are estimated by

$$\vec{\mu}_{A_1} = \frac{1}{2N_t}\sum_{i=1}^{N_t}\left(\vec{\psi}^{\,i}_{A_1B_1} + \vec{\psi}^{\,i}_{A_1B_2}\right) \tag{43}$$

$$\vec{\mu}_{A_2} = \frac{1}{2N_t}\sum_{i=1}^{N_t}\left(\vec{\psi}^{\,i}_{A_2B_1} + \vec{\psi}^{\,i}_{A_2B_2}\right), \tag{44}$$

and C_A is estimated by

$$C_A = \frac{1}{4N_t - 2}\sum_{i=1}^{N_t}\left[\left(\vec{\psi}^{\,i}_{A_1B_1} - \vec{\mu}_{A_1}\right)^2 + \left(\vec{\psi}^{\,i}_{A_1B_2} - \vec{\mu}_{A_1}\right)^2 + \left(\vec{\psi}^{\,i}_{A_2B_1} - \vec{\mu}_{A_2}\right)^2 + \left(\vec{\psi}^{\,i}_{A_2B_2} - \vec{\mu}_{A_2}\right)^2\right], \tag{45}$$

where $\vec{v}^2$ stands for $\vec{v} \cdot \vec{v}^t$.

Next, linear coordinate transform $\tau_1 : \Psi \rightarrow \Psi'$ is carried out on the training data. τ_1 shifts mean $\vec{\mu}_a$ of all training vectors to the origin and "whitens" C_A—that is, maps C_A onto the identity matrix I (Fukunaga, 1990). Matrix T_1 associated with transformation τ_1 contains the eigenvectors $\vec{e}_1$ and $\vec{e}_2$ of covariance matrix C_A, divided by the square root of their respective eigenvalues λ_1 and λ_2:

$$T_1 = \left(\frac{\vec{e}_1}{\sqrt{\lambda_1}}\ \frac{\vec{e}_2}{\sqrt{\lambda_2}}\right). \tag{46}$$

The mapping of an arbitrary vector $\vec{\psi}$ in Ψ onto $\vec{\psi}'$ in Ψ' is thus defined as

$$\vec{\psi}' = T_1^t\left(\vec{\psi} - \vec{\mu}_a\right). \tag{47}$$

Transform 2

τ_2 maps $\vec{\mu}'_{A_1}$ and $\vec{\mu}'_{A_2}$ onto the y-axis (i.e., $\mu''_{A_1x} = \mu''_{A_2x} = 0$), with $\mu''_{A_1y} = \mu''_{A_2y}$, and $\mu''_{A_1y} > 0$, and leaving C'_A unchanged (i.e., $C''_A = I$).

Let C'_a denote the covariance matrix of all τ_1-transformed training vectors. C'_a can be written as

$$C'_a = \frac{1}{2}\left[\left(\vec{\mu}'_{A_1} - \vec{\mu}'_a\right)^2 + \left(\vec{\mu}'_{A_2} - \vec{\mu}'_a\right)^2\right] + C'_A. \tag{48}$$

Because $\vec{\mu}'_a = \vec{0}$ and $C'_A = I$, this reduces to

$$C'_a = \begin{pmatrix} \mu^2_{A_1x} + 1 & \mu_{A_1x}\mu_{A_1y} \\ \mu_{A_1x}\mu_{A_1y} & \mu^2_{A_1y} + 1 \end{pmatrix}. \tag{49}$$

Matrix T_2 associated with coordinate transformation τ_2 is now defined as

$$T_2 = \frac{1}{\sqrt{\mu^2_{A_1x} + \mu^2_{A_1y}}}\begin{pmatrix} \mu_{A_1y} & \mu_{A_1x} \\ -\mu_{A_1x} & \mu_{A_1y} \end{pmatrix}. \tag{50}$$

Note that T_2 contains the eigenvectors of C'_a in *ascending* order of their respective eigenvalues.

The two-step transformation is illustrated for a synthetic data set in Figure 7. The synthetic data set was constructed so that it contains three types of dependencies: shift dependency along the horizontal axis, convergence along the horizontal axis, and convergence along the vertical axis. Panels A, B, and C show the untransformed data set, the data set after τ_1, and the data set after both τ_1 and τ_2, respectively. The ellipses shown in Figure 7 are isoprobability contours of the pdfs associated with categories A_1 and A_2.

Derivation of Optimal Conditional Classifiers

After applying τ_1 and τ_2 to the training data, independence of categorizations A and B, as well as the three types of dependencies of B on A, can be translated into simple conditions on the distributions associated with categories B_1 and B_2 in the double-transformed space Ψ''. These conditions can be derived by applying τ_1 and τ_2 to the idealized pdf geometries of Figure 1, as quantified in the previous section. The conditions thus derived also apply for any linear transformation of these geometries in Ψ, because such linear transformation is removed by τ_1 and τ_2.

Independence. If we apply τ_1 and τ_2 to the *independent* geometry illustrated in Figure 1A and defined in Equations A6–A14, we find the following conditions on the means and covariance matrices of B_1 and B_2:

$$\mu_{B_1y} = \mu_{B_2y} = 0 \tag{51}$$

$$C_{B_1} = C_{B_2} \tag{52}$$

$$\rho_{B_1} = \rho_{B_2} = 0, \tag{53}$$

where ρ_{B_1} is the correlation coefficient of the pdf of B_1. In Equations 51–53 and in the equations below, the dou-

ble primes are left out to improve legibility. Recall that, in addition to Equations 51–53, in general $\vec{\mu}_{B_1} = -\vec{\mu}_{B_2}$, because $\vec{\mu}_a = 0$.

Given a set of phoneme pdfs (i.e., means and covariance matrices for A_1, A_2, B_1, and B_2), we can define *inferred* syllable pdfs (i.e., means and covariance matrices for A_1B_1, A_1B_2, A_2B_1, and A_2B_2) as the syllable pdfs that yield exactly the given phoneme pdfs, with the only restriction being that the covariance matrices of A_1 and A_2 are equal. Application of τ_1 and τ_2 to the geometries of Figure 1 leads to relatively simple relations between the means and covariance matrices of the inferred syllable pdfs and those of the phoneme pdfs. These relations are given in Appendix C.

Position dependence. As compared with the independent model, two conditions on the B_1 and B_2 pdfs are relaxed when the boundary position of categorization B is assumed to depend on A: μ_{B_1y} and μ_{B_2y} are allowed to differ, and ρ_{B_1} and ρ_{B_2} may differ from 0. This leaves only the following condition:

$$C_{B_1} = C_{B_2}. \tag{54}$$

Orientation dependence. Under the assumption that the boundary orientation of categorization B depends on A, the covariance matrices C_{B1} and C_{B2} are no longer equal. Their vertical variances are allowed to differ as follows:

$$\mu_{B_1y} = \mu_{B_2y} = 0 \tag{55}$$

$$\sigma_{B_1x} = \sigma_{B_2x} \tag{56}$$

$$\rho_{B_1} = \rho_{B_2} = 0. \tag{57}$$

Steepness dependence. Under the assumption that the boundary position of categorization B depends on A, again the covariance matrices C_{B1}, C_{B2} are different. This time they have opposite correlation coefficients. The conditions for this situation are given in Equations 58–61:

$$\mu_{B_1y} = \mu_{B_2y} = 0 \tag{58}$$

$$\sigma_{B_1x} = \sigma_{B_2x} \tag{59}$$

$$\sigma_{B_1y} = \sigma_{B_2y} \tag{60}$$

$$\rho_{B_1} = -\rho_{B_2}. \tag{61}$$

The procedures for estimating the optimal conditional classification models are illustrated in Figure 8, using the same set of synthetic data as that for Figure 7.

The first, second, third, and fourth rows of subfigures in Figure 8 illustrate the estimation method for the assumptions of independence, position dependence, orientation dependence, and steepness dependence, respectively. The left-hand and middle columns of subfigures show the double transformed space with the four sets of four data points. In addition to the data points, the subfigures in the left-hand column contain isoprobability contours of the phoneme pdfs for A_1 and A_2 (dashed circles) and for B_1 and B_2 (solid ellipses). The middle column shows, apart from the data, the isoprobability contours of the inferred syllable pdfs (solid ellipses), the A_1–A_2 boundaries (horizontal lines), and the conditional B_1–B_2 boundaries (the more vertically oriented lines). The right-hand column again shows the data, syllable isoprobability contours, and category boundaries, but now in the original space Ψ, after applying the inverse transforms of τ_2 and τ_1 to the data of the middle column. The figure shows how conditions defined by Equations 51–56 on the means and covariance matrices of B_1 and B_2 in Ψ'' are visually expressed in the locations and shapes of the ellipses in the left-hand column and how these translate into syllable ellipses and (conditional) phoneme boundaries in spaces Ψ'' and Ψ.

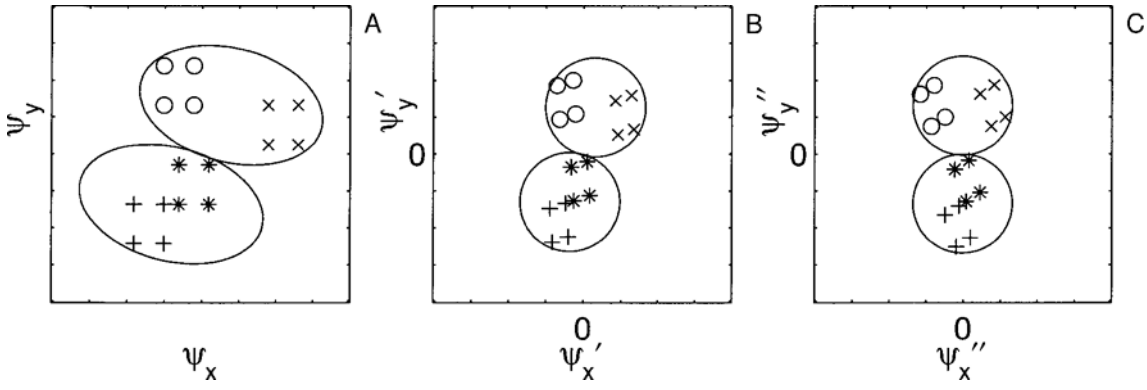

Figure 7. Illustration of the two linear transformations on a synthetic set of acoustic-perceptual data for four syllables (represented by symbols ○, ×, ∗, +). (A) Original data in psychological space Ψ. Ellipses represent isoprobability contours for categories A_1 and A_2, assuming that they have equal covariance matrices. (B) Data in Ψ' after transformation τ_1. Note that as a result of whitening, ellipses for A_1 and A_2 are circles. (C) Data in Ψ'' after transformations τ_1 and τ_2. Note that the means of the circles for A_1 and A_2 are on a vertical line.

Prediction of Dependency Direction and Ordering of Dependency Types

Above, a method was presented for estimating the optimal hierarchical categorization of phonemes, given a set of acoustical measurements. Two questions remain unanswered: How do we predict (1) the most likely dependency *direction* and (2) the most likely dependency *type*? I first address the second point. The method presented below gives a ranking of the usefulness of each of the possible dependency types, given a set of acoustic data. The reasoning behind the method is as follows. The optimal fuzzy classification strategy in a 4RC problem involving two successive phonemes uses the syllable as a unit, as was argued earlier. The most obvious way to evaluate the potential performance gain that is due to the introduction of a dependency in a phoneme-based classifier is, therefore, to compare its fuzzy output to that of the syllable-based classifier. There are two basic ways of doing this: comparing the goodness functions or comparing the probability surfaces. I decided to concentrate on the goodness functions

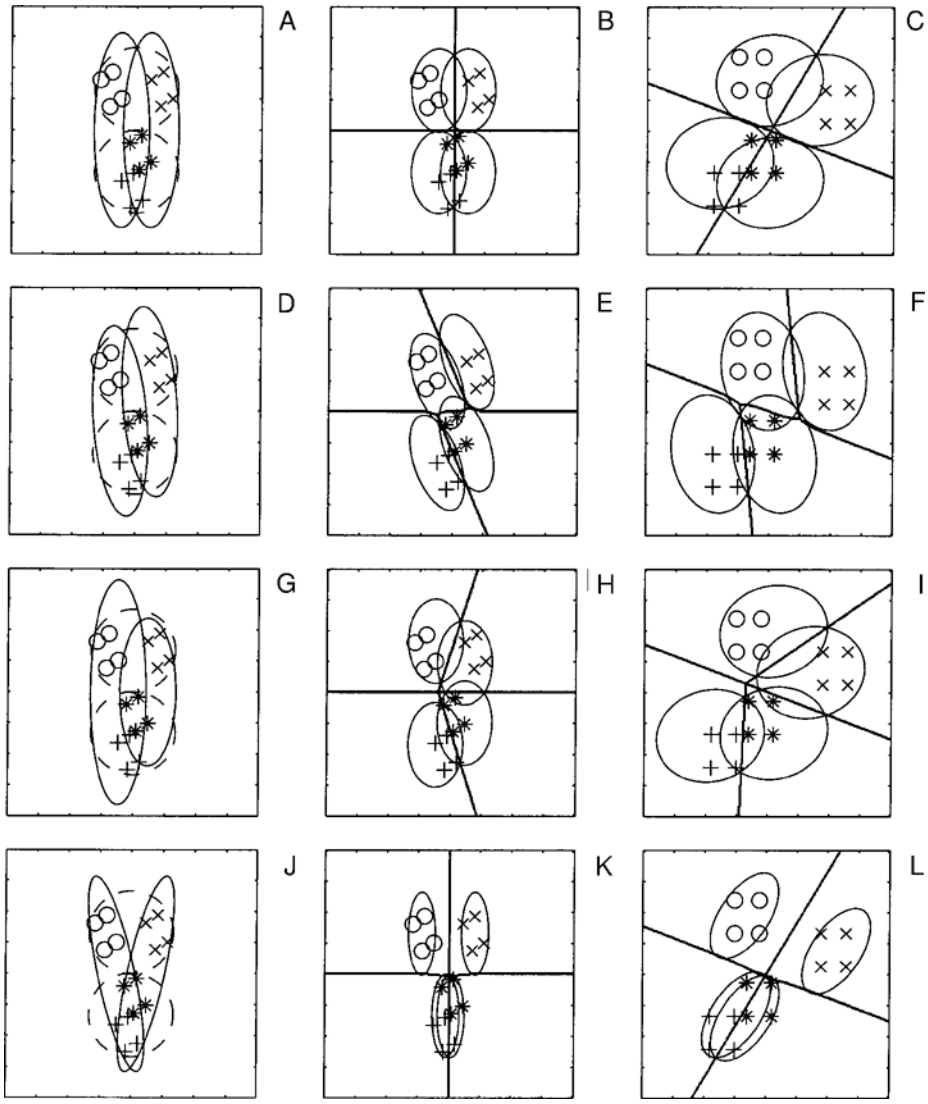

Figure 8. Illustration of the prediction method for the synthetic data set of Figure 7. The first, second, third, and fourth rows of subfigures are associated with assumptions of independence, position dependence, orientation dependence, and steepness dependence, respectively. The left-hand and middle columns show double transformed space Ψ''; the right-hand column shows the original space Ψ. The subfigures in the left-hand column contain isoprobability contours of phoneme pdfs for A_1 and A_2 (dashed circles) and for B_1 and B_2 (solid ellipses). The middle column shows isoprobability contours of inferred syllable pdfs (ellipses), A_1–A_2 boundaries (horizontal lines), and conditional B_1–B_2 boundaries (more vertically oriented lines). The right-hand column shows syllable isoprobability contours and category boundaries in the original space Ψ, after applying inverse transforms of τ_2 and τ_1 to the data of the middle column.

for two reasons. First, it was assumed that the goodness levels, not the probabilities, are actually passed on to higher processing levels during speech recognition. It seems therefore better to base the comparison on goodness functions, rather than on probabilities. The second reason is of a more practical nature. Techniques for comparing Gaussian pdfs are relatively straightforward and are available in the literature. If one wants to compare probability surfaces, on the other hand, one has to make a somewhat arbitrary decision on what section of the space Ψ this comparison should be restricted to. It was therefore decided to express the dissimilarity of the goodness functions corresponding to the inferred syllable pdfs to those corresponding with the actual syllable pdfs in terms of the average Bhattacharyya distance (e.g., Fukunaga, 1990) between corresponding pdfs. Note that this method is compatible with the qualitative definition of classification performance introduced in an earlier section. The calculation therefore results in a ranking of the three dependency types in terms of their performance.

The dependency *direction* most likely to be used by listeners can be predicted by simply selecting the direction that leads to the smallest Bhattacharyya distance. As was discussed earlier, however, an "optimal" dependency direction should really be considered within the larger context of everyday speech perception, where the system has to find a fast but accurate way of dealing with all possible phonemes in all possible contexts within a language. This is a topic for future research.

SUMMARY AND CONCLUSIONS

The present theoretical paper addresses the issue of the perception of coarticulated phonemes. The starting point of the study was the assumption that listeners behave like pattern classifiers that try to deal with the acoustic input in the simplest, but sufficiently accurate, fashion. A reflection on the hypothetical effects of coarticulation on the statistical distributions of acoustic cues produced some useful insights. First, in many cases, the categorization of successive phonemes may perform close to optimally, using a strategy involving an acoustic, but not a phonological context—that is, cues are shared between categorizations, but otherwise the categorizations are independent. Nevertheless, the results suggested that, in cases of severe coarticulation, it might be necessary to adopt a strategy involving some restricted forms of phonological dependency. Three types of hierarchical dependency were proposed: dependency of the position, orientation, and steepness of the category boundary for one categorization on the output of the other categorization.

Next, the HICAT model of hierarchical categorization was defined in which these three types of hierarchical dependency were explicitly modeled. The model involves several processing steps. First, the stimulus is mapped onto a point in multidimensional psychological space. In this space, goodness functions are defined for the relevant phonemes. These functions are Gaussian approximations to the distributions of acoustic cues in the phonemes as they occur in natural speech. A comparison of the stimulus with the relevant goodness functions results in goodness levels for the phonemes. The assumption is made that there is a one-way lateral interaction between the goodness levels of the successive phonemes. Finally, the goodness levels are transformed into probabilities of choosing each of the possible responses, using Luce's choice rule. HICAT's architecture is parallel and bottom-up.

A comparison of HICAT with FLMP revealed a number of differences between the two models. The most important difference was that the 4RC version of FLMP without feature modifiers allows neither acoustic nor phonological context effects, whereas HICAT allows both. FLMP with feature modifiers, on the other hand, was shown to be capable of modeling at least two types of hierarchical dependencies, although in a somewhat different fashion from HICAT. A comparison of HICAT with Nearey's DBSCM showed that the effect of Nearey's diphone bias is similar, although not equal, to the effect of the position dependency parameter in HICAT. HICAT's steepness and orientation dependencies, on the other hand, are absent in DBSCM. A comparison of HICAT with the GRT modeling framework concentrated on perceptual and decisional separability. Essentially, HICAT was shown to allow for violations of both types of independence.

Finally, a method was presented that can be used for predicting, on the basis of a set of measurements of acoustic cue values in natural utterances, which type of dependency is most likely to be used by listeners. This method represents a very strong test of a basic assumption of the pattern-recognition approach, which says that important aspects of listeners' categorization strategies are tuned to the statistical properties of the acoustic *training data*.

REFERENCES

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

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

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

Ashby, F. G., & Maddox, W. T. (1994). A response time theory of separability and integrality in speeded classification. *Journal of Mathematical Psychology*, **38**, 423-466.

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

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

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

Carden, G., Levitt, A., Jusczyk, P. W., & Walley, A. (1981). Evidence for phonetic processing of cues to place of articulation: Perceived manner affects perceived place. *Perception & Psychophysics*, **29**, 26-36.

Cohen, M. M., & Massaro, D. W. (1992). On the similarity of categorization models. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 395-447). Hillsdale, NJ: Erlbaum.

Crowther, C. S., Batchelder, W. H., & Hu, X. (1995). A measurement-theoretic analysis of the fuzzy logic model of perception. *Psychological Review*, **102**, 396-408.

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

Dupoux, E. (1993). The time course of prelexical processing: The syllabic hypothesis revisited. In G. Altmann & R. Shillcock (Eds.), *Cognitive models of speech processing II* (pp. 81-114). Hove, U.K.: Erlbaum.

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

Eimas, P. D., Tartter, V. C., Miller, J. L., & Keuthen, N. J. (1978). Asymmetric dependencies in processing phonetic features. *Perception & Psychophysics*, **23**, 12-20.

Fletcher, H. (1953). *Speech and hearing in communication*. New York: Kreiger.

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.

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

Harris, K. S. (1958). Cues for the discrimination of American English fricatives in spoken syllables. *Language & Speech*, **1**, 1-7.

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

Johnson, K. (1997). The auditory/perceptual basis for speech segmentation. In K. Ainsworth-Darnell & M. D'Imperio (Eds.), *OSU Working Papers in Linguistics* (Vol. 50, pp. 101-113). Columbus: Ohio State University, Department of Linguistics.

Kingston, J., & Macmillan, N. A. (1995). Integrality of nasalization and F_1 in vowels in isolation and before oral and nasal consonants: A detection-theoretic application of the Garner paradigm. *Journal of the Acoustical Society of America*, **97**, 1261-1285.

Kingston, J., Macmillan, N. A., Walsh-Dickey, L., Thorburn, R., & Bartels, C. (1997). Integrality in the perception of tongue root position and voice quality in vowels. *Journal of the Acoustical Society of America*, **101**, 1696-1709.

Kluender, K. R., & Lotto, A. J. (1999). Virtues and perils of an empiricist approach to speech perception. *Journal of the Acoustical Society of America*, **105**, 503-511.

Lahiri, A., & Reetz, H. (1999). The FUL speech recognition system [Abstract]. *Journal of the Acoustical Society of America*, **105**, 1091.

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

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

Lotto, A. J., & Kluender, K. R. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics*, **60**, 602-619.

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

Macmillan, N. A., Goldberg, R. F., & Braida, L. D. (1988). Resolution for speech sounds: Basic sensitivity and context memory on vowel and consonant continua. *Journal of the Acoustical Society of America*, **84**, 1262-1280.

Macmillan, N. A., Kingston, J., Thorburn, R., Walsh-Dickey, L. & Bartels, C. (1999). Integrality of nasalization and F_1: II. Basic sensitivity and phonetic labeling measure distinct sensory and decision-rule interactions. *Journal of the Acoustical Society of America*, **106**, 2913-2932.

Maddieson, I. (1984). *Patterns of sounds.* Cambridge: Cambridge University Press.

Maddox, W. T. (1992). Perceptual and decisional separability. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 147-180). Hillsdale, NJ: Erlbaum.

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

Marslen-Wilson, W. (1987) Functional parallelism in spoken word recognition. In U. H. Frauenfelder & L. K. Tyler (Eds.), *Spoken word recognition* (pp. 71-102). Cambridge, MA: MIT Press.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Erlbaum.

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 context in speech perception. *Perception & Psychophysics*, **34**, 338-348.

Massaro, D. W., & Cohen, M. M. (1991). Integration versus interactive activation: The joint influence of stimulus and context in perception. *Cognitive Psychology*, **23**, 558-614.

Massaro, D. W., & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review*, **97**, 225-252.

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.

Miller, J. L. (1981). Phonetic perception: Evidence for context-dependent and context-independent processing. *Journal of the Acoustical Society of America*, **69**, 822-831.

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

Nearey, T. M. (1992). Context effects in a double-weak theory of speech perception. *Language & Speech*, **35**, 153-171.

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 J. Local (Ed.), *Papers in laboratory phonology VI.* Cambridge: Cambridge University Press.

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

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. 13-44). Orlando, FL: Academic Press.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral & Brain Sciences*, **23**, 299-325.

Oden, G. C. (1992). Direct, incremental learning of fuzzy propositions. In J. K. Kruschke (Ed.), *Proceedings of the 14th Annual Conference of the Cognitive Science Society* (pp. 48-53). Hillsdale, NJ: Erlbaum.

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

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

Repp, B. H. (1980). Accessing phonetic information during perceptual integration of temporally distributed cues. *Journal of Phonetics*, **8**, 185-194.

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

Segui, J., Frauenfelder, U., & Mehler, J. (1981). Phoneme monitoring, syllable monitoring and lexical access. *British Journal of Psychology*, **72**, 471-477.

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

Smits, R. (1997). A pattern-recognition-based framework for research on phonetic perception. In V. Hazan, M. Holland, & S. Rosen (Eds.) *Speech, hearing and language* (Work in Progress 9, pp. 195-229). London: University College London.

SMITS, R. (in press). Evidence for hierarchical categorization of coarticulated phonemes. *Journal of Experimental Psychology: Human Perception & Performance*.

SOLI, S. D. (1981). Second formants in fricatives: Acoustic consequences of fricative–vowel coarticulation. *Journal of the Acoustical Society of America*, **70**, 976-984.

STEVENS, K. N. (1995). Applying phonetic knowledge to lexical access. In J. M. Pardo, E. Enriquez, J. Ortega, J. Ferreiros, J. Macias, & F. J. Valverde (Eds.), *Proceedings Eurospeech 95* (Vol. 1, pp. 3-11). Madrid: Universidad Politécnica Madrid.

SUSSMAN, H. M., MCCAFFREY, H. A., & MATTHEWS, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place of articulation. *Journal of the Acoustical Society of America*, **90**, 1309-1325.

WHALEN, D. H. (1984). Subcategorical phonetic mismatches slow phonetic judgments. *Perception & Psychophysics*, **35**, 49-64.

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.

WOOD, C. C., & DAY, R. S. (1975). Failure of selective attention to phonetic segments in consonant–vowel syllables. *Perception & Psychophysics*, **17**, 346-350.

APPENDIX A
HICAT Derivation From Acoustical Distributions

The following assumptions are made.

1. The unit of recognition is the phoneme.
2. An incoming stimulus S_i is mapped onto a point (ψ_x^i, ψ_y^i) in a two-dimensional psychological space Ψ spanned by axes ψ_x and ψ_y.
3. The category goodness of stimulus S_i for phoneme A_1 is given by a goodness function $\gamma_{A_1}(\psi_x, \psi_y)$ on psychological space Ψ (and analogously for other phonemes).
4. Each goodness function is based on a statistical description of points (ψ_x, ψ_y) associated with previously perceived exemplars of that phoneme. These statistical descriptions are well approximated by multidimensional Gaussian pdfs.
5. For all phonemes, the goodness value at the mean of the distribution (*self-similarity*) equals unity. Thus, the goodness functions are not sensitive to a priori phoneme probabilities.
6. Categorization B depends on categorization A.
7. The Gaussian pdfs associated with phonemes A_1 and A_2 have equal covariance matrices.
8. The Gaussian pdfs associated with syllables A_1B_1, A_1B_2, A_2B_1, and A_2B_2 have equal covariance matrices.
9. The Gaussian pdfs associated with the syllables have covariances equal to zero. (This seems a very strong claim, but it is actually part of the geometry definitions of Figures 1 and 2, as will be discussed later.)

In the derivations presented below, first a geometry of syllable pdfs is defined, and next the optimal hierarchical phoneme classifier for that geometry is derived. During the derivations, I will refer back to Figures 1 and 2, where A is the vowel categorization and B is the fricative categorization. Based on Assumptions 7 and 8, the following relations generally hold between the means $\vec{\mu}_{A_1}$, $\vec{\mu}_{A_2}$ and covariance matrix C_A of phonemes A_1 and A_2 and the means $\vec{\mu}_{A_1B_1}$, $\vec{\mu}_{A_1B_2}$, $\vec{\mu}_{A_2B_1}$, $\vec{\mu}_{A_2B_2}$ and covariance matrix C_{AB} of the four syllables (see, e.g., Fukunaga, 1990):

$$\vec{\mu}_{A_1} = \frac{1}{2}\left(\vec{\mu}_{A_1B_1} + \vec{\mu}_{A_1B_2}\right) \tag{A1}$$

$$\vec{\mu}_{A_2} = \frac{1}{2}\left(\vec{\mu}_{A_2B_1} + \vec{\mu}_{A_2B_2}\right) \tag{A2}$$

$$C_A = C_A^b + C_A^w. \tag{A3}$$

C_A^b and C_A^w in Equation A3 are the between-group and within-group covariance matrices, respectively, defined by

$$C_A^b = \frac{1}{4}\left[\left(\vec{\mu}_{A_1B_1} - \vec{\mu}_{A_1}\right)^2 + \left(\vec{\mu}_{A_1B_2} - \vec{\mu}_{A_1}\right)^2 + \left(\vec{\mu}_{A_2B_1} - \vec{\mu}_{A_2}\right)^2 + \left(\vec{\mu}_{A_2B_1} - \vec{\mu}_{A_2}\right)^2\right] \tag{A4}$$

$$C_A^w = C_{AB}, \tag{A5}$$

where $\vec{v}^2$ is a short notion for $\vec{v} \cdot \vec{v}^t$.

Independence

The geometries of the pdfs of syllable cues in Figure 1A can be defined by the following relations:

$$\mu_{A_1B_1y} = \mu_{A_1B_2y} = \mu_{A_1y} \tag{A6}$$

$$\mu_{A_2B_1y} = \mu_{A_2B_2y} = \mu_{A_2y} \tag{A7}$$

$$\mu_{A_1B_1x} = \mu_{A_2B_1x} = \mu_{B_1x} \tag{A8}$$

$$\mu_{A_1B_2x} = \mu_{A_2B_2x} = \mu_{B_2x} \tag{A9}$$

$$\sigma_{Ax}^2 = \frac{1}{4}\left(\mu_{B_1x} - \mu_{B_2x}\right)^2 + \sigma_{ABx}^2 \tag{A10}$$

APPENDIX A (Continued)

$$\sigma_{Ay}^2 = \sigma_{ABy}^2 \tag{A11}$$

$$\sigma_{Bx}^2 = \sigma_{ABx}^2 \tag{A12}$$

$$\sigma_{By}^2 = \frac{1}{4}\left(\mu_{A_1 y} - \mu_{A_2 y}\right)^2 + \sigma_{ABy}^2 \tag{A13}$$

$$\rho_A = \rho_B = \rho_{AB} = 0. \tag{A14}$$

In general, the goodness function γ_{A_1} is defined as

$$\gamma_{A_1} = \exp - \frac{1}{2\left(1-\rho_A^2\right)}\left[\left(\frac{\psi_x - \mu_{A_1 x}}{\sigma_{Ax}}\right)^2 - 2\rho_A\left(\frac{\psi_x - \mu_{A_1 x}}{\sigma_{Ax}}\right)\left(\frac{\psi_y - \mu_{A_1 y}}{\sigma_{Ay}}\right) + \left(\frac{\psi_y - \mu_{A_1 y}}{\sigma_{Ay}}\right)^2\right], \tag{A15}$$

and analogously for the other phonemes and syllables.

Using Luce's choice rule, the probability $p(A_1)$ of choosing phoneme A_1 is defined as

$$p\left(A_1\right) = \frac{\gamma_{A_1}}{\gamma_{A_1} + \gamma_{A_2}}. \tag{A16}$$

Substituting Equations A6–A15 in Equation A16 and computing the log-odds ratio leads to

$$\ln\left[\frac{p\left(A_1\right)}{p\left(A_2\right)}\right] = \frac{\mu_{A_1 y} - \mu_{A_2 y}}{\sigma_{ABy}^2}\left[\psi_y - \frac{1}{2}\left(\mu_{A_1 y} + \mu_{A_2 y}\right)\right]. \tag{A17}$$

Equation A17 shows that, for the problem given in Figure 1A, the probability of choosing response A_1 (or A_2) as a function of the psychological coordinates is given by a logistic function that depends only on ψ_y. The coefficient of ψ_y—that is,

$$\frac{\mu_{A_1 y} - \mu_{A_2 y}}{\sigma_{ABy}^2}$$

—defines the steepness of the logistic function. Steepness increases with increasing separation between the means and with decreasing variance of γ_{A_1} and γ_{A_2}. The point at which $p(A_1) = p(A_2) = \frac{1}{2}$ gives the boundary location of the logistic function. Here, the boundary location is given by $\psi_y = \frac{1}{2}(\mu_{A1y} + \mu_{A2y})$, which is halfway between the means of γ_{A_1} and γ_{A_2}.

In a similar fashion, the following relation can be derived for the response probabilities for categorization B:

$$\ln\left[\frac{p\left(B_1\right)}{p\left(B_2\right)}\right] = \frac{\mu_{B_1 x} - \mu_{B_2 x}}{\sigma_{ABx}^2}\left[\psi_x - \frac{1}{2}\left(\mu_{B_1 x} + \mu_{B_2 x}\right)\right]. \tag{A18}$$

The probabilities of choosing each of the syllable responses equal the product of the appropriate phoneme probabilities:

$$p\left(A_1 B_1\right) = p\left(A_1\right) \cdot p\left(B_1\right), \tag{A19}$$

and analogously for the other syllables.

Equations A17–A19 are of the same type as Equations 5–7 in the HICAT model definition. It is therefore possible to express the HICAT model parameters in terms of the parameters of the syllable pdfs:

$$p_0 = -\frac{\mu_{A_1 y}^2 - \mu_{A_2 y}^2}{2\sigma_{ABy}^2}, \quad p_x = 0, \quad p_y = \frac{\mu_{A_2 y} - \mu_{A_1 y}}{\sigma_{ABy}^2} \tag{A20}$$

$$q_0 = -\frac{\mu_{B_1 x}^2 - \mu_{B_2 x}^2}{2\sigma_{ABx}^2}, \quad q_x = \frac{\mu_{B_1 x} - \mu_{B_2 x}}{\sigma_{ABx}^2}, \quad q_y = 0 \tag{A21}$$

$$c_0 = c_\alpha = c_\beta = 0. \tag{A22}$$

It is easy to show that, for the geometry of Figure 1A, the syllable probabilities based on the optimal phoneme-based categorization (as given above) are equal to the syllable probabilities based on the optimal syllable-based categorization. A syllable-based categorization strategy is based on syllable goodness functions $\gamma_{A_1B_1}$, $\gamma_{A_1B_2}$, $\gamma_{A_2B_1}$, $\gamma_{A_2B_2}$. The probability of choosing response A_1B_1 is now given by

$$p\left(A_1 B_1\right) = \frac{\gamma_{A_1 B_1}}{\gamma_{A_1 B_1} + \gamma_{A_1 B_2} + \gamma_{A_2 B_1} + \gamma_{A_2 B_2}}, \tag{A23}$$

APPENDIX A (Continued)

and analogously for the other syllables. Substitution of Equations A6–A14 in Equation A23 leads to the same expression as Equation A19.

Boundary Location Dependence

In Figure 1B, the syllable pdfs were shifted along the horizontal axis within a vowel by the amount Δ_x. Means and covariances of the phoneme and syllable pdfs in this geometry are given below:

$$\mu_{A_1B_1x} = \mu_{B_1x} + \Delta_x \tag{A24}$$

$$\mu_{A_2B_1x} = \mu_{B_1x} - \Delta_x \tag{A25}$$

$$\mu_{A_1B_2x} = \mu_{B_2x} + \Delta_x \tag{A26}$$

$$\mu_{A_2B_2x} = \mu_{B_2x} - \Delta_x \tag{A27}$$

$$\mu_{A_1B_1y} = \mu_{A_1B_2y} = \mu_{A_1y} \tag{A28}$$

$$\mu_{A_2B_1y} = \mu_{A_2B_2y} = \mu_{A_2y} \tag{A29}$$

$$\sigma^2_{Ax} = \frac{1}{4}\left(\mu_{B_1x} - \mu_{B_2x}\right)^2 + \sigma^2_{ABx} \tag{A30}$$

$$\sigma^2_{Ay} = \sigma^2_{ABy} \tag{A31}$$

$$\rho_A = \rho_{AB} = 0. \tag{A32}$$

Using Equations A15 and A16, we find

$$\ln\left[\frac{p(A_1)}{p(A_2)}\right] = \frac{2\Delta_x}{\sigma^2_{Ax}}\left[\psi_x - \frac{1}{2}\left(\mu_{B_1x} + \mu_{B_2x}\right)\right] + \frac{\mu_{A_1y} - \mu_{A_2y}}{\sigma^2_{ABy}}\left[\psi_y - \frac{1}{2}\left(\mu_{A_1y} + \mu_{A_2y}\right)\right]. \tag{A33}$$

The optimal *conditional* categorization for B given A_1 is, by definition, based on the syllable goodness functions γ_{A1B1} and γ_{A1B2}:

$$p(B_1 \mid A_1) = \frac{\gamma_{A_1B_1}}{\gamma_{A_1B_1} + \gamma_{A_1B_2}}. \tag{A34}$$

This leads to

$$\ln\left[\frac{p(B_1 \mid A_1)}{p(B_2 \mid A_1)}\right] = \frac{\mu_{B_1x} - \mu_{B_2x}}{\sigma^2_{ABx}}\left(\psi_x - \frac{1}{2}\left(\mu_{B_1x} + \mu_{B_2x}\right) - \Delta_x\right). \tag{A35}$$

Equation A35 shows that that the optimal boundary between B_1 and B_2 given A_1 is identical to the optimal B_1–B_2 boundary in the independence case shifted to the right by the amount Δ_x. The optimal conditional categorization for B given A_2 leads to an expression for

$$\ln\left[\frac{p(B_1 \mid A_2)}{p(B_2 \mid A_2)}\right]$$

that is identical to Equation A35, with Δ_x replaced by $-\Delta_x$—that is, a boundary shift of Δ_x to the left.

Expressions A33 and A35 are of the same type as Expressions 5–7 in the HICAT model definition. The relations of HICAT model parameters to parameters of the syllable pdfs are as follows:

$$p_0 = -\frac{\Delta_x\left(\mu_{B_1x} + \mu_{B_2x}\right)}{\frac{1}{4}\left(\mu_{B_1x} - \mu_{B_2x}\right)^2 + \sigma^2_{ABx}} - \frac{\mu^2_{A_1y} - \mu^2_{A_2y}}{2\sigma^2_{ABy}},$$

$$p_x = \frac{2\Delta_x}{\frac{1}{4}\left(\mu_{B_1x} - \mu_{B_2x}\right)^2 + \sigma^2_{ABx}}, \quad p_y = \frac{\mu_{A_1y} - \mu_{A_2y}}{\sigma^2_{ABy}} \tag{A36}$$

$$q_0 = -\frac{\mu^2_{B_1x} - \mu^2_{B_2x}}{2\sigma^2_{ABx}}, \quad q_x = \frac{\mu_{B_1x} - \mu_{B_2x}}{\sigma^2_{ABx}}, \quad q_y = 0 \tag{A37}$$

$$c_0 = -\Delta_x \frac{\mu_{B_1x} - \mu_{B_2x}}{\sigma^2_{ABx}}, \quad c_\alpha = c_\beta = 0. \tag{A38}$$

Boundary Orientation Dependence

Figure 2C showed two different strategies for dealing with the *converged* geometry of Figure 1C. The dashed lines of Figure 2C indicate the strategy with the orientation dependence, but here the vowel catego-

APPENDIX A (Continued)

rization is dependent on the fricative. For an orientation dependence with the reverse dependency, we consider the geometry of Figure 2C rotated 90° counterclockwise—that is, with a convergence of the pdfs within a fricative category along the vertical axis. Means and covariances of the phoneme and syllable pdfs are given below:

$$\mu_{A_1B_1x} = \mu_{A_2B_1x} = \mu_{B_1x} \tag{A39}$$

$$\mu_{A_1B_2x} = \mu_{A_2B_2x} = \mu_{B_2x} \tag{A40}$$

$$\mu_{A_1B_1y} = \mu_{A_1y} - \Delta_y \tag{A41}$$

$$\mu_{A_1B_2y} = \mu_{A_1y} + \Delta_y \tag{A42}$$

$$\mu_{A_2B_1y} = \mu_{A_2y} + \Delta_y \tag{A43}$$

$$\mu_{A_2B_2y} = \mu_{A_2y} - \Delta_y \tag{A44}$$

$$\sigma^2_{Ax} = \frac{1}{4}\left(\mu_{B_1x} - \mu_{B_2x}\right)^2 + \sigma^2_{ABx} \tag{A45}$$

$$\sigma^2_{Ay} = \Delta^2_y + \sigma^2_{ABy} \tag{A46}$$

$$\rho_A = \rho_{AB} = 0. \tag{A47}$$

Using Equations A15 and A16, we find that

$$\ln\left[\frac{p(A_1)}{p(A_2)}\right] = \frac{\mu_{A_1y} - \mu_{A_2y}}{\Delta^2_y + \sigma^2_{ABy}}\left[\psi_y - \frac{1}{2}\left(\mu_{A_1y} + \mu_{A_2y}\right)\right]. \tag{A48}$$

The optimal *conditional* categorization for B given A_1 is again based on the syllable goodness functions $\gamma_{A_1B_1}$ and $\gamma_{A_1B_2}$, which leads to

$$\ln\left[\frac{p(B_1 \mid A_1)}{p(B_2 \mid A_1)}\right] = \frac{\mu_{B_1x} - \mu_{B_2x}}{\sigma^2_{ABx}}\left[\psi_x - \frac{1}{2}\left(\mu_{B_1x} + \mu_{B_2x}\right)\right] - \frac{2\Delta_y}{\sigma^2_{ABy}}\left(\psi_y - \mu_{A_1y}\right). \tag{A49}$$

Equation A49 shows that the optimal boundary between B_1 and B_2 given A_1 is identical to the optimal B_1 and B_2 boundary in the independence case rotated clockwise around the point

$$\left[\frac{1}{2}\left(\mu_{B_1x} + \mu_{B_2x}\right),\ \mu_{A_1y}\right]$$

The optimal conditional categorization for B given A_2 leads to an expression for

$$\ln\left[\frac{p(B_1 \mid A_2)}{p(B_2 \mid A_2)}\right]$$

that is identical to Equation A49, with Δ_y replaced by $-\Delta_y$ and μ_{A_1y} replaced by μ_{A_2y}.

The relations of HICAT model parameters to parameters of the syllable pdfs are as follows:

$$p_0 = -\frac{\mu^2_{A_1y} - \mu^2_{A_2y}}{2\left(\sigma^2_{ABy} + \Delta^2_y\right)},\ p_x = 0,\ p_y = \frac{\mu_{A_1y} - \mu_{A_2y}}{\sigma^2_{ABy} + \Delta^2_y}, \tag{A50}$$

$$q_0 = \frac{\Delta_y\left(\mu_{A_1y} - \mu_{A_2y}\right)}{\sigma^2_{ABy}} - \frac{\mu^2_{B_1x} - \mu^2_{B_2x}}{2\sigma^2_{ABx}},\quad q_x = \frac{\mu_{B_1x} - \mu_{B_2x}}{\sigma^2_{ABx}},\quad q_y = 0 \tag{A51}$$

$$c_\alpha = -\frac{2\Delta_y}{\sigma^2_{ABy}}, c_0 = c_\beta = 0. \tag{A52}$$

Boundary Steepness Dependence

The dotted lines in Figure 2C showed the categorization strategy with the steepness dependence. Equations A53–A61 give the means and covariances of the phoneme and syllable pdfs for this case:

$$\mu_{A_1B_1x} = \mu_{B_1x} + \Delta_x \tag{A53}$$

$$\mu_{A_2B_1x} = \mu_{B_1x} - \Delta_x \tag{A54}$$

$$\mu_{A_1B_2x} = \mu_{B_2x} - \Delta_x \tag{A55}$$

$$\mu_{A_2B_2x} = \mu_{B_2x} + \Delta_x \tag{A56}$$

APPENDIX A (Continued)

$$\mu_{A_1B_1y} = \mu_{A_1B_2y} = \mu_{A_1y} \tag{A57}$$

$$\mu_{A_2B_1y} = \mu_{A_2B_2y} = \mu_{A_2y} \tag{A58}$$

$$\sigma_{Ax}^2 = \frac{1}{4}\left(\mu_{B_1x} - \mu_{B_2x}\right)^2 + \Delta_x^2 + \sigma_{ABx}^2 \tag{A59}$$

$$\sigma_{Ay}^2 = \sigma_{ABy}^2 \tag{A60}$$

$$\rho_A = \rho_{AB} = 0. \tag{A61}$$

Using Equations A15 and A16, we find that

$$\ln\left[\frac{p(A_1)}{p(A_2)}\right] = \frac{\mu_{A_1y} - \mu_{A_2y}}{\sigma_{ABy}^2}\left[\psi_y - \frac{1}{2}\left(\mu_{A_1y} + \mu_{A_2y}\right)\right]. \tag{A62}$$

The optimal *conditional* categorization for B given A_1 is again based on the syllable goodness functions γ_{A1B1} and γ_{A1B2}, which leads to

$$\ln\left[\frac{p(B_1 \mid A_1)}{p(B_2 \mid A_1)}\right] = \frac{\mu_{B_1x} - \mu_{B_2x} + 2\Delta_x}{\sigma_{ABx}^2}\left[\psi_x - \frac{1}{2}\left(\mu_{B_1x} + \mu_{B_2x}\right)\right]. \tag{A63}$$

Equation A63 shows that that the optimal boundary between B_1 and B_2 given A_1 is identical to the optimal B_1–B_2 boundary in the independence case but is steeper. The expression for

$$\ln\left[\frac{p(B_1 \mid A_2)}{p(B_2 \mid A_2)}\right]$$

is equal to Equation A63, with Δ_y replaced by $-\Delta_y$—that is, here the boundary is shallower than that for the independent case.

The relations of HICAT model parameters to parameters of the syllable pdfs are as follows:

$$p_0 = -\frac{\mu_{A_1y}^2 - \mu_{A_2y}^2}{2\sigma_{ABy}^2},\ p_x = 0,\ p_y = \frac{\mu_{A_1y} - \mu_{A_2y}}{\sigma_{ABy}^2}, \tag{A64}$$

$$q_0 = -\frac{\mu_{B_1x}^2 - \mu_{B_2x}^2}{2\sigma_{ABx}^2},\quad q_x = \frac{\mu_{B_1x} - \mu_{B_2x}}{\sigma_{ABx}^2},\quad q_y = 0 \tag{A65}$$

$$c_\beta = \frac{2\Delta_x}{\sigma_{ABx}^2}, c_0 = c_\alpha = 0. \tag{A66}$$

APPENDIX B
Mathematics of Parallel HICAT Implementation

In this appendix, it is shown that the adjustment of the goodness functions γ_{B1} and γ_{B2} in the parallel architecture as described earlier leads to the correct HICAT expressions. The strategy is as follows. First, we express the conditional (i.e., adjusted) goodness functions for B as general functions of the unconditional goodness functions for A and B:

$$\gamma_{B_1|A_1} = \gamma_{B_1}\left(\frac{\gamma_{B_1}}{\gamma_{B_2}}\right)^{p_1+p_2}\left(\frac{\gamma_{A_1}}{\gamma_{A_2}}\right)^{q_1+q_2} e^{r_1+r_2} \tag{B1}$$

$$\gamma_{B_2|A_1} = \gamma_{B_2}\left(\frac{\gamma_{B_1}}{\gamma_{B_2}}\right)^{-p_1-p_2}\left(\frac{\gamma_{A_1}}{\gamma_{A_2}}\right)^{-q_1-q_2} e^{-r_1-r_2} \tag{B2}$$

$$\gamma_{B_1|A_2} = \gamma_{B_1}\left(\frac{\gamma_{B_1}}{\gamma_{B_2}}\right)^{-p_1+p_2}\left(\frac{\gamma_{A_1}}{\gamma_{A_2}}\right)^{-q_1+q_2} e^{-r_1+r_2} \tag{B3}$$

$$\gamma_{B_2|A_2} = \gamma_{B_2}\left(\frac{\gamma_{B_1}}{\gamma_{B_2}}\right)^{p_1-p_2}\left(\frac{\gamma_{A_1}}{\gamma_{A_2}}\right)^{q_1-q_2} e^{r_1-r_2}. \tag{B4}$$

APPENDIX B (Continued)

Next, Luce's choice rule is applied to these conditional goodness functions:

$$p(B_1 \mid A_1) = \frac{\gamma_{B_1|A_1}}{\gamma_{B_1|A_1} + \gamma_{B_2|A_1}} \tag{B5}$$

$$p(B_1 \mid A_2) = \frac{\gamma_{B_1|A_2}}{\gamma_{B_1|A_2} + \gamma_{B_2|A_2}}. \tag{B6}$$

If the conditional probabilities $p(B_1 \mid A_1)$ and $p(B_1 \mid A_2)$ in Equations B5 and B6 are optimal, they are by definition equal to $p(B_1 \mid A_1)$ and $p(B_1 \mid A_2)$ derived from the syllable probabilities:

$$p(B_1 \mid A_1) = \frac{\gamma_{A_1B_1}}{\gamma_{A_1B_1} + \gamma_{A_1B_2}} \tag{B7}$$

$$p(B_1 \mid A_2) = \frac{\gamma_{A_2B_1}}{\gamma_{A_2B_1} + \gamma_{A_2B_2}}. \tag{B8}$$

Substitution of Equations B1–B4 in Equations B5 and B6 and application of Equations B7 and B8 leads to a set of equations expressing factors containing the exponents p_1, p_2, q_1, q_2, r_1, and r_2 in terms of γ_{B1}, γ_{B2}, and γ_{A1B1}, γ_{A1B2}, γ_{A2B1}, γ_{A2B2}.

Next, these expressions are applied to the pdf geometries of Figures 1 and 2, using parameter definitions Equations A24–A61. Tedious but straightforward calculations lead to the following relations between exponents p_1, p_2, q_1, q_2, r_1, r_2 and the parameters of the pdf geometries in the three dependency situations.

Position Dependency

$$p_1 = q_1 = r_2 = 0 \tag{B9}$$

$$p_2 = \frac{\Delta_x^2 \sigma_{ABy}^2 \left(\mu_{B_1x} - \mu_{B_2x}\right)^2}{8\sigma_{ABx}^2 \left\{\left[\frac{1}{4}\left(\mu_{A_1y} - \mu_{A_2y}\right)^2 + \sigma_{ABy}^2\right]\left[\frac{1}{4}\left(\mu_{B_1x} - \mu_{B_2x}\right)^2 + \sigma_{ABx}^2\right] + \Delta_x^2 \sigma_{ABy}^2\right.} \tag{B10}$$

$$q_2 = \frac{2\,\Delta_x \sigma_{ABy}^2 \left(\mu_{B_1x} - \mu_{B_2x}\right)\left[\frac{1}{4}\left(\mu_{B_1x} - \mu_{B_2x}\right)^2 + \sigma_{ABx}^2\right]}{8\sigma_{ABx}^2 \left\{\left[\frac{1}{4}\left(\mu_{A_1y} - \mu_{A_2y}\right)^2 + \sigma_{ABy}^2\right]\left[\frac{1}{4}\left(\mu_{B_1x} - \mu_{B_2x}\right)^2 + \sigma_{ABx}^2\right] + \Delta_x^2 \sigma_{ABy}^2\right\}} \tag{B11}$$

$$r_1 = -\frac{\Delta_x\left(\mu_{B_1x} - \mu_{B_2x}\right)}{2\sigma_{ABx}^2}. \tag{B12}$$

Orientation Dependency

$$p_1 = p_2 = q_2 = r_1 = 0 \tag{B13}$$

$$q_1 = -\frac{\Delta_y\left(\Delta_y^2 + \sigma_{ABy}^2\right)}{\sigma_{ABy}^2\left(\mu_{A_1y} - \mu_{A_2y}\right)} \tag{B14}$$

$$r_2 = \frac{\Delta_y\left(\mu_{A_1y} - \mu_{A_2y}\right)}{2\sigma_{ABy}^2}. \tag{B15}$$

Steepness Dependency

$$p_2 = q_1 = q_2 = r_1 = r_2 = 0 \tag{B16}$$

$$p_1 = \frac{\Delta_x^2\left(\mu_{B_1x} - \mu_{B_2x}\right) + 2\Delta_x^3 + 2\Delta_x\sigma_{ABx}^2}{2\sigma_{ABx}^2\left(\mu_{B_1x} - \mu_{B_2x}\right)}. \tag{B17}$$

APPENDIX C
Relations between Inferred Syllable Pdfs and Phoneme Pdfs in Double-Transformed Space

This appendix lists the relations between parameters of the inferred syllable distributions and the phoneme distributions in the double-transformed space Ψ. The double primes are left out to improve legibility.

Independence

$$\mu_{A_1B_1x} = \mu_{A_2B_1x} = \mu_{B_1x} = -\mu_{A_1B_2x} = -\mu_{A_2B_2x} = -\mu_{B_2x} \tag{C1}$$

$$\mu_{A_1B_1y} = \mu_{A_1B_2y} = \mu_{A_1y} = -\mu_{A_2B_1y} = -\mu_{A_2B_2y} = -\mu_{A_2y} \tag{C2}$$

$$\sigma_{ABx} = \sigma_{Bx} \tag{C3}$$

$$\sigma_{ABy} = \sigma_{Ay} = 1 \tag{C4}$$

$$\rho_{AB} = 0. \tag{C5}$$

Position Dependence

$$\mu_{A_1B_1x} = \mu_{A_2B_1x} = \mu_{B_1x} = -\mu_{A_1B_2x} = -\mu_{A_2B_2x} = -\mu_{B_2x} \tag{C6}$$

$$\mu_{A_1B_1y} = \mu_{A_1y} + \mu_{B_1y} = -\mu_{A_2B_2y} = -\mu_{A_2y} - \mu_{B_2y} \tag{C7}$$

$$\mu_{A_2B_1y} = \mu_{A_2y} + \mu_{B_1y} = -\mu_{A_1B_2y} = -\mu_{A_1y} - \mu_{B_2y} \tag{C8}$$

$$\sigma_{ABx} = \sigma_{Bx} \tag{C9}$$

$$\sigma^2_{ABy} = \sigma^2_{By} - \frac{1}{4}\left(\mu_{A_1y} - \mu_{A_2y}\right)^2 \tag{C10}$$

$$\sigma_{ABx}\sigma_{ABy}\rho_{AB} = \sigma_{Bx}\sigma_{By}\rho_B. \tag{C11}$$

Orientation Dependence

$$\mu_{A_1B_1x} = \mu_{A_2B_1x} = \mu_{B_1x} = -\mu_{A_1B_2x} = -\mu_{A_2B_2x} = -\mu_{B_2x} \tag{C12}$$

$$\mu_{A_1B_1y} = \mu_{A_1y} - \Delta_y = -\mu_{A_2B_1y} = -\mu_{A_2y} + \Delta_y \tag{C13}$$

$$\mu_{A_1B_2y} = \mu_{A_1y} + \Delta_y = -\mu_{A_2B_2y} = -\mu_{A_2y} - \Delta_y \tag{C14}$$

$$\Delta_y = \frac{\sigma^2_{B_2y} - \sigma^2_{B_1y}}{2\left(\mu_{A_1y} - \mu_{A_2y}\right)} \tag{C15}$$

$$\sigma_{ABx} = \sigma_{Bx} \tag{C16}$$

$$\sigma^2_{ABy} = \frac{1}{2}\left(\sigma^2_{B_1y} + \sigma^2_{B_2y}\right) - \frac{1}{4}\left(\mu_{A_1y} - \mu_{A_2y}\right)^2 - \Delta^2_y \tag{C17}$$

$$\rho_{AB} = 0. \tag{C18}$$

Steepness Dependence

$$\mu_{A_1B_1y} = \mu_{A_1B_2y} = \mu_{A_1y} = -\mu_{A_2B_1y} = -\mu_{A_2B_2y} = -\mu_{A_2y} \tag{C19}$$

$$\mu_{A_1B_1x} = \mu_{B_1x} + \Delta_x = -\mu_{A_1B_2x} = -\mu_{B_2x} - \Delta_x \tag{C20}$$

$$\mu_{A_2B_1x} = \mu_{B_1x} - \Delta_x = -\mu_{A_2B_2x} = -\mu_{B_2x} + \Delta_x \tag{C21}$$

$$\Delta_x = \frac{\sigma_{Bx}\sigma_{By}\left(\rho_{B_1} - \rho_{B_2}\right)}{\mu_{A_1y} - \mu_{A_2y}} \tag{C22}$$

$$\sigma^2_{ABx} = \sigma^2_{Bx} - \Delta^2_x \tag{C23}$$

$$\sigma^2_{ABy} = \sigma^2_{Ay} = 1 \tag{C24}$$

$$\rho_{AB} = 0. \tag{C25}$$

(Manuscript received March 28, 2000;
revision accepted for publication January 15, 2001.)