Cognitive Science 34 (2010) 434–464
Copyright © 2009 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2009.01077.x

Cue Integration With Categories: Weighting Acoustic Cues in Speech Using Unsupervised Learning and Distributional Statistics

Joseph C. Toscano, Bob McMurray

Department of Psychology and Delta Center, University of Iowa

Received 24 November 2008; received in revised form 15 May 2009; accepted 3 August 2009

Abstract

During speech perception, listeners make judgments about the phonological category of sounds by taking advantage of multiple acoustic cues for each phonological contrast. Perceptual experiments have shown that listeners weight these cues differently. How do listeners weight and combine acoustic cues to arrive at an overall estimate of the category for a speech sound? Here, we present several simulations using a mixture of Gaussians models that learn cue weights and combine cues on the basis of their distributional statistics. We show that a cue-weighting metric in which cues receive weight as a function of their reliability at distinguishing phonological categories provides a good fit to the perceptual data obtained from human listeners, but only when these weights emerge through the dynamics of learning. These results suggest that cue weights can be readily extracted from the speech signal through unsupervised learning processes.

Keywords: Speech perception; Speech development; Cue weighting; Reliability; Categorization; Statistical learning; Unsupervised learning; Mixture of Gaussians

1. Introduction

In every domain of perception, multiple sources of information must be combined. A classic example is depth perception, where the distance of an object from an observer is indicated by a number of cues, including stereopsis, perspective, binocular disparity, shading, motion, and many others (see Kaufman, 1974, for an extensive list). We use the term *cue* here to refer to any source of information that allows the perceiver to distinguish

Correspondence should be sent to Joseph Toscano, Department of Psychology, E11 SSH, University of Iowa, Iowa City, IA 52242. E-mail: joseph-toscano@uiowa.edu

between different responses. Each cue provides a continuous estimate of depth, and to get an accurate estimate, observers must combine information across them.

This raises the question of how much weight or importance should be assigned to each cue. An emerging consensus is that cues are weighed as a function of the reliability of the estimates they provide (Ernst & Banks, 2002; Jacobs, 1999; Landy & Kojima, 2001) and that cue reliability can be learned (Atkins, Fiser, & Jacobs, 2001). Some depth cues, like stereopsis, provide robust estimates (Johnson, Cummings, & Landy, 1994), whereas other cues, like shading are relatively poor (Bülthoff & Mallot, 1988). Weighting cues based on their reliability (Jacobs, 2002; Kalman, 1960) offers a formal approach for estimating these weights and using them to arrive at a combined estimate. Using this method, the weight of an individual cue at a specific depth is determined by:

$$w = \frac{1}{\sigma^2} \tag{1}$$

where w is the weight of the cue and σ^2 is the variance of the estimate provided by that cue at a given depth (i.e., how accurately that cue allows the observer to estimate depth). The overall depth estimate, X, can then be calculated as a linear combination of the weighted cue estimates:

$$X = \sum_{i}^{n} w_i x_i \tag{2}$$

This approach has been shown to be consistent with observers' performance in a number of tasks (Battaglia, Jacobs, & Aslin, 2003; Ernst & Banks, 2002; Jacobs, 1999).

Weighting-by-reliability works well when cue integration can be described as the linear combination of continuous cues and when their variance is roughly Gaussian. However, for many perceptual problems, the causal factors that give rise to the cues are not themselves continuous. In these cases, the perceptual system faces the joint problem of recovering both a continuous estimate of the perceptual cue and also the underlying categories that shaped it.

Speech perception provides an excellent example of this. In speech, phonological dimensions like voicing (which distinguishes voiced sounds like /b, d, g/ from voiceless sounds like /p, t, k/) are often determined by a large number of continuous acoustic cues. For example, cues to word-initial voicing include voice onset time (VOT; Liberman, Harris, Kinney, & Lane, 1961), vowel length (VL; Miller & Dexter, 1988; Summerfield, 1981), pitch (Haggard, Ambler, & Callow, 1970), and F1 onset frequency (Stevens & Klatt, 1974). Understanding how listeners combine these cues, often described behaviorally using trading relations (a shift in the identification function for one cue with changes in another cue), is central to understanding speech perception (see Repp, 1982, for a review of trading relations in speech).

While these cues are continuous, their statistical distributions are shaped into clusters of cue values by the phonological categories of the language. The listener's goal is to determine the underlying phonological category from these cues, not necessarily a continuous estimate (although there is evidence that listeners also estimate continuous values and the likelihood of a category; see Massaro & Cohen, 1983; McMurray, Tanenhaus, & Aslin, 2002; Schouten, Gerrits, & van Hessen, 2003). Thus, the goal for speech perception is slightly different than the goal for depth perception. In depth perception, observers must recover the best estimate of the depth (i.e., determine a quantity along a metric dimension), whereas the goal in speech perception is to recover the best estimate of a discrete underlying category.

The presence of categories makes it difficult to apply the weighting-by-reliability approach directly, as variance along the cue dimension itself does not map onto how well that cue supports categorization. For example, if we look at the frequency distribution of values for VOT in Fig. 1A, there are clusters of cue values corresponding to voiced sounds (VOTs near 0 ms) and voiceless sounds (VOTs near 50 ms). This clustering makes a simple computation of reliability from the variance in the estimator of a cue impractical. However, it also enables us to compute a different metric of reliability. That is, the relevant variance

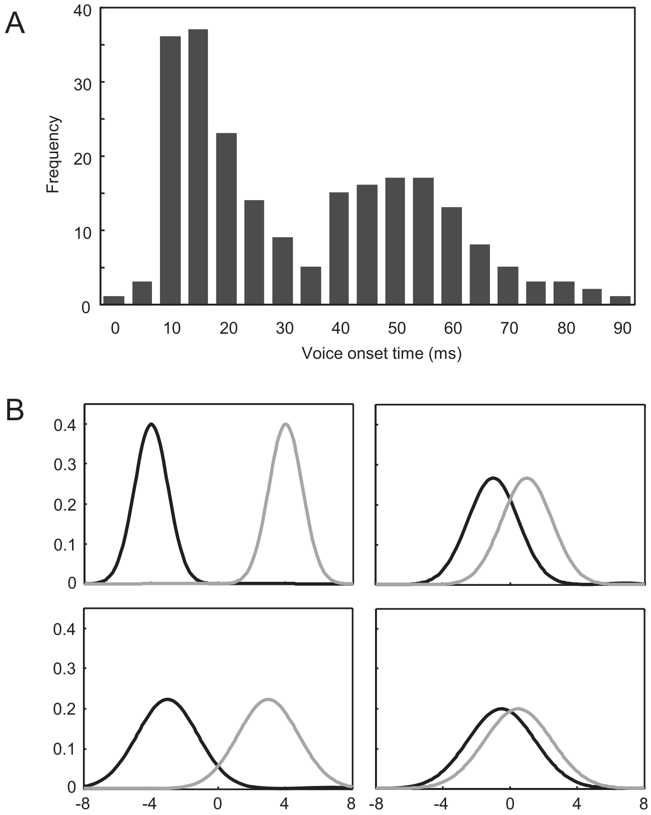

Fig. 1. (A) Distribution of VOT values for voiced and voiceless stops in English (from acoustic measurements in Allen & Miller, 1999). (B) Two-category cues can vary in how reliable they are (and, in turn, how highly they should be weighted) from reliable cues with distinct categories (top left) to unreliable cues with overlapping categories (bottom right). Cues of intermediate reliability (top right and bottom left) require us to take into account both the distance between categories and the variability within them.

for speech perception is the variance in the ability of the cue to support categorization. In a sense, the variance of VOT as an estimator of voicing is a function of both the variability within the categories and the distance between them. Thus, the reliability of a cue or dimension can only be determined with respect to the underlying phonological categories.

Given that the distributional statistics of cues must be learned to compute weights based on their reliability, it is fundamental to take into account what is known about development. Thus, whatever solution we adopt to the problem of cue weighting should be based on knowledge and representations that are developmentally plausible. In particular, as we will describe below, the acquisition of speech discrimination abilities may derive from a similar estimation of distributional statistics. The goal of the present work is to adapt the weighting-by-reliability approach in a way that is consistent with the process of speech development.

1.1. Models of cue integration in speech

Formal approaches to cue integration in speech preceded the weighting-by-reliability approach used in depth perception and thus do not incorporate an explicit notion of reliability in their solutions to this problem. The fuzzy logical model of perception (FLMP; Massaro & Oden, 1980; Oden & Massaro, 1978) provides one of the best formal approaches and has successfully modeled a range of cue-integration problems. In its mappings between dimensions and categories, it is clear that some notion of weighting emerges. However, it does not completely solve the problem of cue weighting for two reasons. First, it assigns independent weights to different regions of the same dimensions (e.g., VOTs between 0 and 10 may get substantial weight, but VOTs between 10 and 20 may get less), creating a sparse data problem (i.e., how does a listener deal with a new value along a familiar dimension?). Second, and more importantly, the weights are fit to perceptual data, rather than estimated from the structure of the speech input. Listeners' responses in speech tasks certainly reflect their own cue weights; thus, these weights will be implicitly incorporated into FLMPs integration rules. However, this does not provide an explanation for why listeners would weight one cue over another and does not allow us predict perceptual data from acoustic measurements alone.

Nearey and colleague's normal a posteriori probability (NAPP) models offer a similar approach to cue integration (Nearey, 1997; Nearey & Assmann, 1986; Nearey & Hogan, 1986;). NAPP models use discriminant analysis to assign tokens to categories based on a set of acoustic cues. Like FLMP, these classifications can be probabilistic, allowing the output of the model to be compared with listeners' identification rates. However, unlike FLMP, NAPP models use measurements from production data along with the intended categories to classify tokens. As in FLMP, the training categories (i.e., the intended production) capture some of the differential variability between dimensions, suggesting that NAPP models may also show implicit weighting effects.

Both NAPP models and FLMP treat cue integration as a category-dependent process. To weight cue dimensions in this way, listeners would have to know which tokens belong to which category. However, since category membership is not available to listeners from the acoustic input, the problem of acquiring categories, as well as learning cue weights, might be better characterized as an unsupervised clustering process (Maye, Werker, & Gerken, 2002; McMurray, Aslin, & Toscano, 2009a). Infants tune their phonological discrimination abilities to the native language well before any words are known (e.g., Werker & Curtin, 2005; Werker & Tees, 1984 for a review), and, thus, the development of speech categories must be at least partially category independent. Given this, it makes sense to seek a category-independent way to describe cue integration and weighting that is sensitive to this unsupervised developmental process.

1.2. Weighting cues in speech

The weighting-by-reliability approach and NAPP models appear to offer some insights for solving the problem of cue weighting, as they allow us to estimate cue weights independently of perception. While, as stated above, the cue-weighting method used in depth perception is not adequate for acoustic cues in speech, it does offer some intuitions about how to proceed. In addition, NAPP models suggest that the distributional statistics of acoustic cues can provide the information needed to weight them. Thus, combining the strengths of these two approaches may yield a more complete model.

Fig. 1B shows several possible categories imposed on a given dimension. The top-left panel shows a dimension that would appear reliable: The categories are far apart and have low variability. Conversely, if categories are close together and have high within-category variability (bottom right panel) this dimension should receive little weight. For more ambiguous cases (top right and bottom left), determining reliability is a function of both the distance between the categories and within-category variability, weighed by their respective variances. In support of this, Clayards, Tanenhaus, Aslin, and Jacobs (2008) demonstrated that artificially manipulating the variance of an acoustic cue changes how listeners weight it perceptually.

These intuitions can be captured formally, by treating each cluster as an independent Gaussian distribution. In this case, we can partial out the overall variance along a dimension into the component due to the difference between category means and the variance within each category. This leads to a simple way to estimate the reliability of a dimension:

$$w = \frac{(\mu_1 - \mu_2)^2}{\sigma_1 \sigma_2} \tag{3}$$

Here, μ_1 and μ_2 are the means of each category (e.g., /b/ and /p/), and σ_1 and σ_2 are their standard deviations. This metric would provide listeners with an estimate of cue reliability that is similar to the one provided by the weighting-by-reliability method used in vision. It is similar to standard statistical measures that compare the variance between groups $(\mu_1 - \mu_2)^2$, to the variance within groups, $\sigma_1\sigma_2$. When both σs are equal, this is a pairwise F-ratio.

This solution requires that listeners are sensitive to the distributional statistics of acoustic cues and that cue weights can be based on and learned from this information.

There is growing consensus that listeners are sensitive to these statistics and, further, that infants use statistical learning mechanisms to acquire speech sound categories. Maye and colleagues, for example, have demonstrated that after a brief exposure to statistically structured input, infants discriminate speech sounds consistent with the number of clusters along dimensions like VOT (Maye et al., 2002; Maye, Weiss, & Aslin, 2008; see also Teinonen, Aslin, Alku, & Csibra, 2008). Thus, at the coarsest level of analysis, listeners are likely to have access to and can learn from the statistics necessary for cue weighting.

However, attempts to implement this approach computationally suggest that statistical category learning is not trivial (e.g., de Boer & Kuhl, 2003; McMurray et al., 2009a). In particular, when the number of categories is not known (as languages can carve up the same dimensions in many ways) and the input is not tagged with the underlying category, there is no analytic solution to the problem of estimating the parameters that describe the means, variances, and frequencies of the speech categories. Thus, while at a first approximation, our intuitive modification of the weighting-by-reliability approach seems reasonable, it is significantly underdeveloped from the perspective of learnablity.

The purpose of the present work is to bridge this gap and adapt the weighting-by-reliability approach to the problem of cue integration in speech. Recently, McMurray et al. (2009a) presented a mixture of Gaussians (MOG) model that solves many of the problems of unsupervised learning of phonological categories. This model offers a computational-level description (Marr, 1982) of speech sound categorization while also including a mechanistic account of the developmental process. Here, we extend this model to multiple dimensions and demonstrate how the weighting-by-reliability approach can be implemented in it. In doing so, we reveal some surprising findings about the role of learning processes in statistical cue weighting and the role of context in shifting apparent cue weights.

2. Model architectures

2.1. Mixture of Gaussians models of speech categories

A distribution of acoustic cues can be described as a mixture of probability distributions, in which the likelihood of a given cue value (x) is the product of two factors: (1) the prior probability of each category and (2) the conditional probability of x given each category. This latter probability is usually described as a continuous distribution of cue values, given the parameters of that category. In typical instantiations, for a particular category, the values of a particular cue cluster around the category mean in a Gaussian distribution (although other distributions are possible).

A number of recent studies have modeled the distribution of speech cues using this framework (e.g., de Boer & Kuhl, 2003; McMurray et al., 2009a; Vallabha, McClelland, Pons, Werker, & Amano, 2007). In general, cues in these models are represented by a set of

Gaussian distributions (Fig. 2A) each defined by three parameters: frequency of occurrence (ϕ), mean (μ), and standard deviation (σ) (Fig. 2B). Thus, the likelihood of a particular cue-value (x) for each Gaussian is:

$$G_i(x) = \phi_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right) \tag{4}$$

and the overall likelihood is the sum of the likelihoods for each Gaussian:

$$M(x) = \sum_i^K G_i(x) \tag{5}$$

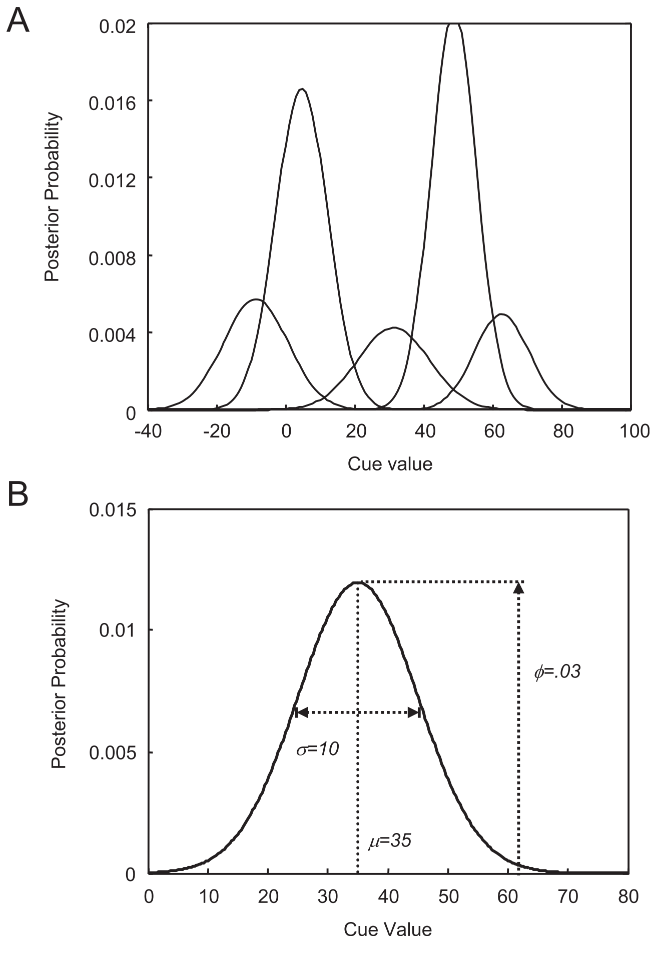

Fig. 2. (A) A mixture of Gaussians with five categories along the dimension. (B) Parameters of a Gaussian distribution used in the mixture model. Each distribution is defined by three parameters—its likelihood (ϕ), its mean (μ), and its standard deviation (σ).

Here, K represents the number of Gaussians in the mixture. For example, the likelihood of a VOT of 30 ms is the sum of the relatively high probability that it arose from a /p/ ($\mu \approx 50$, $\sigma \approx 15$) and the lower probability it came from a /b/ ($\mu \approx 0$, $\sigma \approx 5$).

The fact that this model explicitly represents parameters like μ and σ makes it an ideal platform for implementing the variation of the weighting-by-reliability approach described in Eq. 3. However, it also raises the critical question of how to determine the mixture's parameter values. One option would be to simply set the values of K, ϕ, μ, and σ for each category using values extracted from acoustic measurements. However, there is no analytic solution to parameter estimation in a mixture model when the underlying categories for each data point are unknown, and expectation maximization and similar learning algorithms have a difficult time estimating the number of categories in this situation (de Boer & Kuhl, 2003). Simply assigning values on the basis of such measurements, then, assumes that listeners have access to some learning mechanism that is not guaranteed to exist.

McMurray et al. (2009a) demonstrated how gradient descent can be used to model the gradual acquisition of speech categories when learning is unsupervised and the number of categories is not known. Their model uses large values of K (e.g., 20), with the expectation that ϕs will be reduced to near zero for unneeded categories. The idea here is that, over training, K itself does not change, but, because most Gaussians will have very small values for ϕ, the mixture will functionally behave as if K was only 2 (for a two-category dataset). Parameters are updated via maximum likelihood estimation using the derivatives of the probability density function with respect to each parameter (see Appendix for these learning rules). A crucial innovation from this model is the use of a winner-take-all update rule for ϕ such that only one Gaussian updates its ϕ value on any given trial. This allows the model to suppress unneeded categories and arrive at the correct solution. Without it, the model does not determine the correct number of categories (see McMurray et al., 2009a). This winner-take-all competition is similar to competitive learning approaches used in other unsupervised category learning models like SUSTAIN (Love, Medin, & Gureckis, 2003) as well as various neural network models (McMurray, Horst, Toscano, & Samuelson, 2009b; Rumelhart & Zipser, 1985). McMurray et al. (2009a) describe how this simple solution allowed 97/100 models in their simulations to arrive at the correct two-category solution for a corpus of VOT measurements, and they demonstrate that the time course of learning shows many parallels with infant speech development.

This implementation raises two issues for the cue-weighting approach we have described. First, as there are more than two categories (even if only two will be used eventually), our weighting metric must be able to handle many categories and factor out unused ones. Second, because this model is based on gradient descent, the learning procedure is not guaranteed to find the globally optimal parameter values based on the distributional statistics of the data (i.e., it may settle in a local minimum). This raises the question of whether the dynamics of learning affect cue weighting. If learning leads to nonoptimal representations in the model that reflect human behavior, it would suggest that listeners may be behaving in a way that is not entirely consistent with the statistics of the input. We examine both of these issues in our simulations.

2.2. Cue integration in a mixture of Gaussians

The MOG framework allows us to incorporate underlying categories into the reliability estimates described above because it explicitly represents those categories, allowing us to relate the distance between μs to the corresponding σs. The cue-weighting strategy we have discussed would allow us to combine estimates from different cues into a single overall estimate whose inputs are weighted by the reliability of the individual cues. We also consider, as a comparison, an alternative approach in which multiple cues are represented in a multidimensional MOG with individual cues along separate dimensions (e.g., two-dimensional Gaussians for two cues). In this case, weighting emerges implicitly. (This is not necessarily a criticism of the model; indeed, this property is useful as it allows us to model cue integration without having to specify an additional function for determining cue weights.) This multidimensional model, which is highly parameterized and can represent distributions completely, serves as a baseline model for comparison with a more constrained cue-weighting model.

2.2.1. Cue-weighting model

To compute a cue weight, we must know the variability in the estimate for a given cue-value. Eq. 3 describes one way to do this in the specific case of a two-category cue. However, many phonological contrasts contain more than two categories (e.g., voicing in Thai; place of articulation in English). In addition, using the unsupervised learning approach from McMurray et al. (2009a), the model will have many more possible categories than the number of categories in the data. Thus, at some points during learning, it may not be possible to know which Gaussian will become a particular adult category.

Thus, we use a weighting metric that captures a more general case in which the cue dimension may have any number of categories. In our model, the weight of an individual acoustic cue (i) is:

$$w_i = \left(\sum_{n}^{K} \sum_{m}^{K} \frac{\phi_m \phi_n (\mu_m - \mu_n)^2}{\sigma_m \sigma_n} \right) / 2 \tag{6}$$

This metric is similar to Eq. 3. However, it allows for any number of categories by summing all of the pairwise comparisons between the parameters of the Gaussians in the mixture (i.e., each pair of Gaussians, m and n, from 1 to K) and then halving this sum so that each pair does not contribute twice to the weight.

Two features are worth noting. First, pairs of Gaussians whose means are far apart will increase the weight of the cue, but this is balanced by the within-category variability of those Gaussians. If within-category standard deviations are large, the weight of the cue will be smaller. Thus, as with Eq. 3, this metric is similar to measures like d' or the t statistic, in which both the distance between group means and within-group variances are taken into account.

Second, although K is large, most of the Gaussians are unused after training. This would seem to clutter up the computation with unnecessary comparisons. However, as unused Gaussians will have ϕs near zero, they will not contribute much to the weight. This allows us to compute the weight of a cue regardless of the number of categories and without knowing which specific Gaussians correspond to each category.

After computing the weight for each cue, the weighted estimates are combined to obtain a continuous overall estimate along an underlying phonological dimension. This overall estimate serves as input to an additional MOG that represents the abstract phonological feature distinguished by the cues (e.g., voicing). Thus, the cue-level MOGs are used only to compute weights and inputs to the combined MOG—category judgments are made on the basis of the combined MOG itself. Note that, similar to the cues themselves, this combined MOG is based on a continuous phonological representation. However, it is abstracted away from the input (as it represents a combination of cues). Thus, it is similar to other proposals that phonological representations are continuous (e.g., Frisch, 1996), but it stands in contrast to models that have proposed a more direct mapping between input and phonology, such as exemplar models (Goldinger, 1998; Pierrehumbert, 2001, 2003).

Because different cues give estimates measured on different scales, the combined MOG cannot be based directly on the raw values for each cue. Thus, cue values are normalized by converting the inputs for individual cues to z-scores using the grand mean and variance of each dimension. In addition, the particular ordering of categories along each cue dimension may not be the same for all cues. For example, in the specific case of the two acoustic cues studied here (VOT and VL), voiced sounds are associated with *short* VOTs but *long* VLs. Thus, z-scores for VL are multiplied by the sign of the raw correlation between these two cues across all categories ($r = -0.196$ for VOT/VL data from Allen & Miller, 1999) to deal with differences in the relative ordering of categories along each dimension. Finally, the normalized estimates are then weighted (per Eq. 6) and summed. Fig. 3 shows a schematic representation of the cue weighting and combination process in the model.

2.2.2. Multidimensional model

The cue-weighting model can be contrasted with a model that represents categories in a higher-dimensional acoustic space. In this model, categories are multidimensional Gaussians, and each cue lies along a separate dimension. Thus, for two cues, categories would be represented by bivariate Gaussian distributions (Fig. 4; Eq. A1). This allows the model to take advantage of the entire acoustic space and does not require it to explicitly weight cues. Cue-weighting can emerge implicitly when the categories along one dimension are wide and overlapping, while the other is narrow.

This approach raises several problems. First, the number of parameters in the model can be quite large. For a set of categories determined by a large number of cues (which is not uncommon; see Jongman, Wayland, & Wong, 2000; Lisker, 1986), the model would have to estimate a large number of parameters for each category (e.g., for 16 cues [the number reported by Lisker, 1986], the model would have to estimate 168 parameters for each category). In contrast, in the cue-weighting model, only three

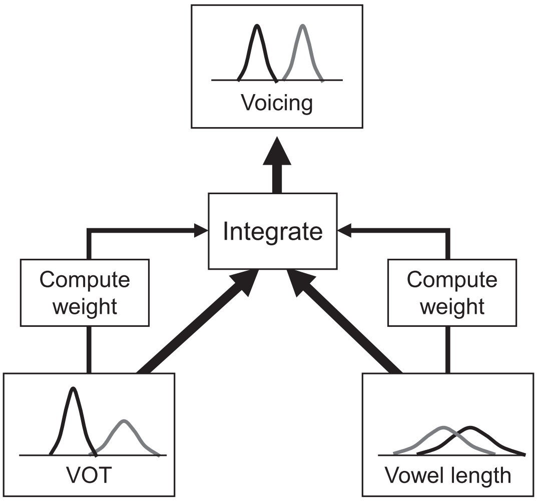

Fig. 3. Schematic representation of the cue-weighting model. Two MOGs receive input for each cue (VOT and VL). The cue weight for each MOG is computed using Eq. 6, and the inputs are converted to *z*-scores. The inputs are then weighted and summed, providing input to a third MOG that reflects voicing categories based on both cues.

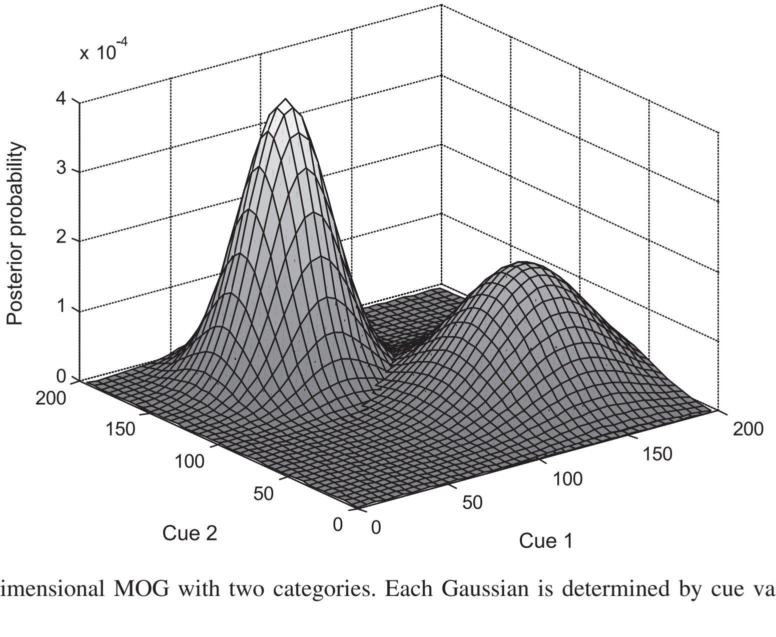

Fig. 4. A two-dimensional MOG with two categories. Each Gaussian is determined by cue values along both dimensions.

parameters per category need to be estimated (for each individual dimension), along with an additional set for the combined dimension (e.g., for 16 cues, the model would only need 51 parameters for each category).

Second, since a given input is a point in a high-dimensional space, sampling may be sparse and many regions of acoustic space will never be encountered during training. This makes it difficult to estimate all of the parameters accurately. The cue-weighting model in contrast considers each dimension independently (not in combination) and may have less trouble with this. Thus, the cue-weighting model may be preferred, if it is better able to learn the categories for a larger set of cues.

3. Simulations

3.1. Acoustic and behavioral data

We ran simulations with each of these models using two acoustic cues to word-initial voicing in English: VOT, mentioned above, and VL. While VL is a robust cue to word-*final* voicing in English (Peterson & Lehiste, 1960; Warren & Marslen-Wilson, 1987), the length of the vowel following the consonantal release has long been recognized as a weak cue to word-*initial* voicing (Allen & Miller, 1999; Miller & Dexter, 1988; Miller & Volaitis, 1989; Summerfield, 1981).

Typically, longer vowels are produced for voiced stops and shorter vowels for voiceless stops. For example, Fig. 5A shows a scatter plot of the VOT and VL values from Allen and Miller (1999). Along the VOT dimension, the categories are highly distinct, reflecting the fact that VOT is a strong cue to voicing. Along the VL dimension, the categories are distinguishable, but highly overlapping, suggesting that VL is a weaker cue. Our training data were similar. VOTs were randomly generated from the means reported by Lisker and Abramson (1964) and VLs from the data in Allen and Miller (1999) (see Table 1).

Empirical work has demonstrated trading relations between these cues: The category boundary along a VOT continuum shifts for different VLs (McMurray, Clayards, Tanenhaus, & Aslin, 2008; Miller & Volaitis, 1989; Summerfield, 1981). Near the boundary, stimuli with long VLs are more often categorized as voiced, and short VLs are more often labeled voiceless (see Fig. 5B for representative results from McMurray et al., 2008). While previous approaches suggested that this trading relation might be an instance of speaking rate compensation (Summerfield, 1981), later work has distinguished it from sentential rate (Wayland, Miller, & Volaitis, 1994; see Repp, 1982 for a discussion of evidence that VL is distinct from overall speaking rate).

Whatever way we characterize the effect of VL, identification functions like the one in Fig. 5B suggest that listeners use both cues but rely more heavily on VOT, an effect that mirrors the statistical distributions of each cue. Simulations 1 and 2 examined whether the multidimensional and cue-weighting models also show a similar trading relation between VOT and VL.

3.2. Simulation 1: VOT and VL in the multidimensional model

The first simulation provides a baseline for performance on the VOT/VL task. Fifty repetitions of two-dimensional MOGs were trained using data sampled from the VOT and VL distributions described above.[1] Initial μ values were randomly chosen from a distribution

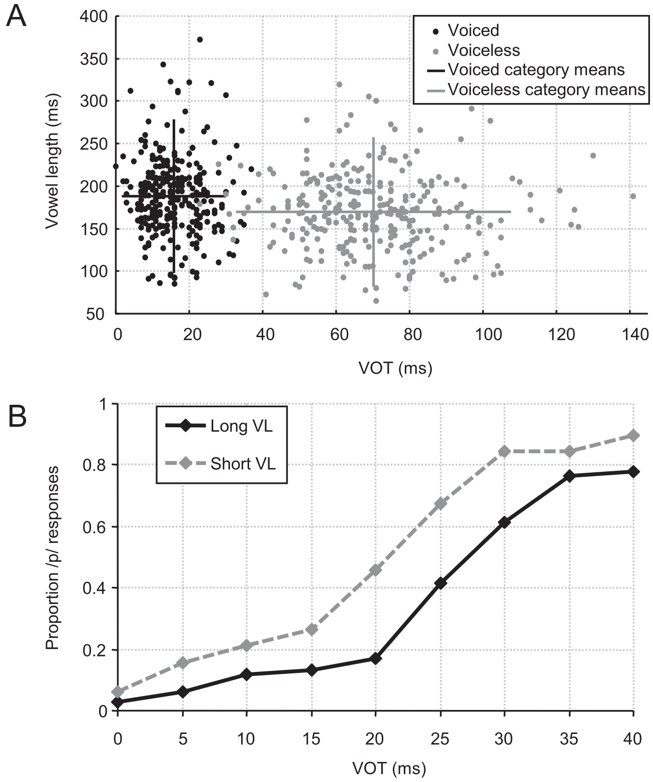

Fig. 5. (A) Distributions of VOT and VL values for voiced and voiceless stops from the production data in Allen and Miller (1999). The locations of the lines indicate the mean of each category along each dimension, and the lengths of the lines are equal to the standard deviations of each distribution. Voiced and voiceless sounds are primarily distinguished along the VOT dimension, but there is also a small difference along the VL dimension. (B) Identification responses from listeners for sounds varying in both VOT and VL from McMurray et al. (2008). Listeners tend to identify sounds as voiced (/b/) for low VOTs and voiceless (/p/) for long VOTs. The shift in the identification function for the two different VL conditions indicates that they also use VL information when making voicing judgments (i.e., they are more likely to identify sounds as voiced for long VLs and voiceless for short VLs).

with a mean of 25 and standard deviation of 75 for the VOT dimension and a mean of 179 and standard deviation of 75 for the VL dimension. Initial σs were set to 3 for the VOT dimiension and 10 for the VL dimension. K was set to 20, and initial ϕs were set to $1/K$. Learning rates were set to 1 (η_μ), 1 (η_σ), 0.001 (η_ϕ), and 0.001 (η_ρ).[2] The models were then tested on a range of VOTs (0–40 ms in 5 ms steps) and two VLs (125 and 225 ms).[3]

3.2.1. Procedure

Each model was trained on 200,000 data points. On each trial, a pair of VOT/VL values was selected for input, and the parameters of the Gaussians in the mixture were updated via the gradient descent learning algorithm discussed above. Winner-take-all competition was implemented by selecting the Gaussian with the highest posterior

Table 1
Descriptive statistics for the distributions used to generate training data

	Voiced			Voiceless		
	VOT (ms)	VL (ms)	Third Cue	VOT (ms)	VL (ms)	Third Cue
Mean	0	188	260	50	170	300
SD	5	45	10	10	44	10

Note. VOT and VL cues were used in Simulations 1–4. The ‘‘third cue’’ is an additional cue that was used in Simulation 4.

probability for that input and updating only that Gaussian's ϕ parameter. The ϕs were then normalized so that they summed to 1. Thus, for the winning Gaussian, ϕ increased, and the others decreased slightly. Only Gaussians with a ϕ value above a threshold of 0.1 were analyzed. Typically, the model arrived at a solution with two above-threshold Gaussians (i.e., the voiced and voiceless categories) with ϕ values of $\approx$0.5. Models that overgeneralized (i.e., arrived at a one-category solution) or did not have any above-threshold Gaussians at the end of training were excluded from analysis. While this seems like a relatively coarse way to assess model performance, we found that if ϕ falls below a threshold of about 0.1, that Gaussian does not typically recover (other Gaussians ultimately represent the two categories). Moreover, in an analysis of the one-dimensional model, we found that if the model arrived at two categories, μ and σ were almost always accurate.[4]

After training, the model was tested using a procedure similar to the task used in McMurray et al. (2008). The model was presented with a pair of VOT and VL values, and identification responses were computed from the posterior probability for each category, which were then normalized using the Luce choice rule (Luce, 1959; temperature = 1) to obtain the proportion of /p/ responses.

3.2.2. Results and discussion

The model learned this distribution quite well, with every repetition adopting the two-category solution. The parameter estimates were also close to the values in the dataset; the average deviation of μ from the category mean was 1.7 ms for VOT and 5.9 ms for VL (see Table 2).

Fig. 6 shows the mean proportion of /p/ responses for the model. A clear effect of VOT is observed, with short VOTs producing more /b/ responses and long VOTs producing more /p/ responses. In contrast, the effect of VL is absent, unlike the effect observed in the empirical data. This was not because of a failure of learning—the model learned the distributions of the two cues and correctly determined the number of categories. In fact, on average the model reported means of 1.63 and 51.3 for VOT, and 188 and 178 for VL, suggesting that it had closely captured the statistics of the input (compare to the means of the training distributions in Table 1). This, however, led to a much weaker trading relation than was observed behaviorally. Since the categories along the VL dimension are highly

Table 2
Mean parameter values at the end of training for Simulations 1, 2, and 4.

		Voiced			Voiceless		
		VOT	VL	Third Cue	VOT	VL	Third Cue
Simulation 1	μ	1.37	188	–	51.5	178	
	σ	5.58	44.8	–	10.2	44.8	–
	ϕ	0.493	0.493	–	0.491	0.491	–
	ρ	−0.067	−0.067	–	−0.040	−0.040	–
Simulation 2	μ	0.253	203	–	50.1	154	–
	σ	5.41	38.4	–	10.0	38.4	–
	ϕ	0.499	0.502	–	0.501	0.498	–
Simulation 4	μ	−0.067	204	260	50.0	155	300
	σ	5.35	38.7	10.1	10.1	38.3 .	10.2
	ϕ	0.501	0.498	0.502	0.499	0.502	0.498

Note. Values for the third cue are only applicable to Simulation 4. There is only a single ϕ and ρ for each voicing category in Simulation 1, since there is only a single value for each of these parameters in the two-dimensional Gaussians.

overlapping, the model relied on VOT instead of VL. Thus, changes along the VOT dimension produced large changes in the model's identification of voicing category, while changes along the VL dimension did not affect the model's category judgments. This result does not reflect listeners' behavior, suggesting that listeners may actually assign more weight to VL than they should be based solely on the statistics of the input.

3.3. Simulation 2: VOT and VL integration in the cue-weighting model

We now consider the cue-weighting model, which consists of three one-dimensional MOGs. The first two represent the VOT and VL dimensions and were used to compute cue

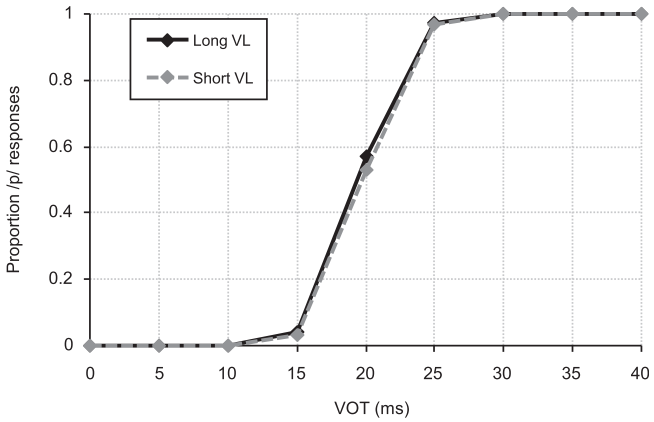

Fig. 6. Identification results for the multidimensional model from Simulation 1. A clear effect of VOT is observed, but the expected trading relation (i.e., the effect of VL) is not.

weights. The third MOG represents categories based on the combination of the two cues (i.e., a voicing dimension) and is used to compute the actual phonological judgments. Fifty repetitions were trained on data sampled from the same distributions as in Simulation 1. Learning rates, K, and the initial σs and ϕs were the same as in Simulation 1; the initial μ values were chosen in the same way.

3.3.1. Procedure

As in Simulation 1, parameters were updated using gradient descent learning and winner-take-all competition. Each model was trained on 90,000 data points, and models that overgeneralized were excluded from analysis. On each trial during training, the parameters of the Gaussians in the VOT and VL MOGs were updated. Then weights were computed for each of these cue-level MOGs using Eq. 6. Next, the input values for the combined MOG were computed by converting the inputs for the individual cues to z-scores, negating the sign for the VL input (because of the negative correlation between the cues), weighting the inputs, and summing them according to Eq. 2. The update procedure was then repeated for the Gaussians in the combined mixture. The testing procedure was the same as the one used for the multidimensional model, except that the posteriors for each category were computed from the combined MOG.

3.3.2. Results and discussion

Overall, this model performed similarly to the prior model with 46/50 models showing the correct two-category solution. Of the ones that failed, three overgeneralized (a single category in one of the MOGs) and one did not have any above-threshold categories in the combined MOG.

Fig. 7 shows the responses of the model in the categorization task. A moderate trading relation was observed, similar to the behavioral data from McMurray et al. (2008). The average cue weight for the VOT dimension was 0.95, and the average cue weight for VL

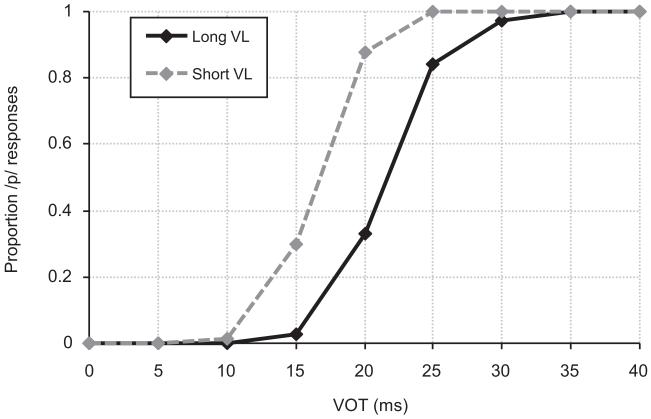

Fig. 7. Identification results for the cue-weighting model from Simulation 2. An effect of both VOT and VL is observed, consistent with responses from listeners.

was 0.05. The results of this simulation suggest that the cue-weighting model shows a good fit to the empirical results, demonstrating that cue weights can be learned from the distributional statistics of the acoustic cues in the input and that a full multidimensional model is not necessary for combining acoustic cues along a single phonological dimension.

Why did the cue-weighting model show a trading relation while the multidimensional model did not? As mentioned-above, the multidimensional model correctly fit the statistics of the dataset. In the cue-weighting model, however, the categories for the VL dimension were further apart than the means in the dataset (mean for each category in the model: 203 [/b/] and 154 [/p/] ms; means in the dataset: 188 [/b/] and 170 [/p/] ms). This caused the model to give more weight to VL, resulting in a trading relation.

This exaggeration of the VL categories may have been a result of the fact that the categories along the VL dimension are highly overlapping. Since training is unsupervised, this exaggeration cannot be because of a performance benefit for representing categories in this way (although there may be one). Thus, during learning there may have been a local minimum that was more stable than the actual means and variances in the data. Indeed, if we were to compute the cue weights using Eq. 6 directly from the acoustic data, we would obtain a relative weight of 0.997 for VOT and 0.003 for VL and, consequently, a significantly reduced trading relation (Fig. 8). This much more closely reflects the behavior of the multidimensional model, not the cue-weighting model or listeners. Thus, the cue-weighting model may have produced different results because its parameters and weights were the product of learning, not derived veridically from the input.

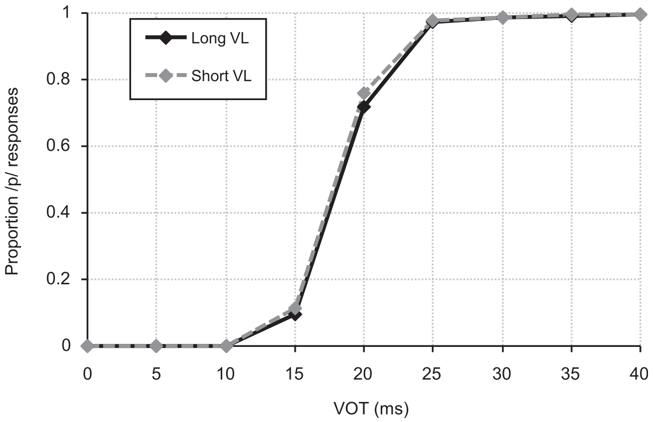

Fig. 8. Predicted results for the cue-weighting model based on the distributional statistics of the two cues. These responses were obtained by setting the parameters of the Gaussians in the input-level MOGs to the means and standard deviations of each cue and training the model (i.e., keeping the parameters at the cue-level MOGs constant while allowing the model to learn the parameters at the combined MOG). The model shows a much smaller trading relation than the one observed in Simulation 2 and in the data from listeners.

3.4. Simulation 3: Effects of learning on VL cue weight

To test the hypothesis that the overweighting of the VL dimension in the cue-weighting model was the product of learning, we ran an additional simulation in which K, the number of Gaussians in the mixture, was set to 2. While this value is quite a bit smaller than the one used in the initial simulations, it allows us to manually set the initial values for μ to observe their behavior over the course of learning. Thus, the starting μs were set to different points along the VL dimension (closer together than the category means, exactly equal to them, or further apart) to determine what the learning algorithm would do in each case. All other parameters were the same as those used in Simulation 2. This allows us to ask what outcomes the learning rules impose on the models' representation of VL beyond those determined by the statistics of the input. For example, if the model started with the correct μ-values and the categories were still forced apart, this would suggest that the category means are not a stable point in state space. Fifty repetitions of this simulation were run. Initial σ and ϕ values were the same as those in the first two simulations.

3.4.1. Results and discussion

The proportion of successful models, as measured by whether the model arrived at a two-category (successful) or one-category (unsuccessful) solution, is shown in Fig. 9B. None of the models with starting μs between the two category means succeeded. Furthermore, every model whose μ values started *outside* the observed local minima values was successful. Thus, the model needed to start with μs that were further apart than necessary.[5]

Fig. 9A shows the change in μs over time. For all models, μs were initially pushed apart (even if that pushed them beyond their eventual location) and over time evolved to the values observed in Simulation 2. Thus, the dynamics of learning seem to favor this exaggeration along the VL dimension.

These results suggest that there are attractor points that the model arrives at through learning in which category means are further apart than the means in the data. Thus, learning (instantiated here by our gradient descent update rules) may be critical for determining the weight of individual cues. Indeed, without learning, the cue-weighting model does not reflect the responses from human listeners. With learning, the model only succeeds under conditions in which it learns categories that are more distinct than those in the data. This suggests that human listeners may not behave in a way that optimally reflects the statistics of the input and that this may result from the fact that speech sound categories are acquired in part through an unsupervised process.

3.5. Simulation 4: Covarying cues in cue-weighting model

So far, we have assumed that the magnitude of trading relations directly reflects the relative weight of the cues. However, previous work on cue integration in speech has shown that the presence of other cues can influence trading relations. For example, Shinn, Blumstein, and Jongman (1985) examined listeners' use of VL and for-

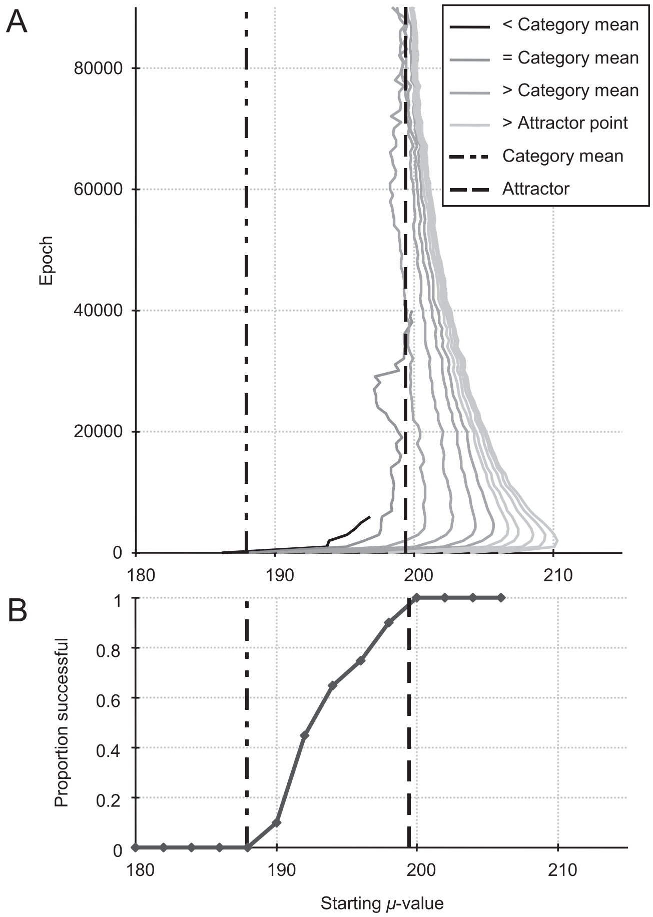

Fig. 9. Results of Simulation 3 for the voiced VL category. The bottom panel shows that when the starting μ values of the Gaussians were initially closer than the means in the data, the model always failed to learn a two-category solution (i.e., it overgeneralized the dimension into a single category). As the starting values were moved further apart, the model became more likely to succeed. The top panel shows the μ values over the course of learning for the different starting values. The point at which each line ends indicates the latest point during training at which one of the models for that starting value still maintained a two-category solution. All of the models that had two categories throughout training settled on μ values near points further apart than the means in the dataset.

mant transition duration (TD) on manner (/b/-/w/) distinctions. In addition to these cues, they simultaneously manipulated additional cues that covaried with TD to produce more natural continua. In this case, the trading relation between TD and VL was reduced. The additional covarying cues may have caused listeners to ignore VL (although see Miller & Wayland, 1993). Utman (1998) found a similar effect, showing that the trading relation between VOT and VL is reduced in natural speech, which contains a large number of voicing cues (see also Lisker, 1975).

Under the assumption that trading relations are determined largely by the relative weight of the available cues, these results imply that listeners perceive synthetic and natural speech differently, reweighting cues depending on the type of input. However, differences in the observed trading relations may reflect other factors besides the weight of the cues. For example, if changes in a third cue are correlated with changes in one of the other two cues, the relative weight of the correlated cues may appear larger—together they are effectively more reliable.

This is straightforward to test in the cue-weighting model. The model can be tested using different stimuli without changing the weights to examine whether additional cues have an effect on trading relations. Thus, we trained the model with a third, artificial cue and tested it under conditions in which, during testing, this cue either covaried with VOT or was held constant at an ambiguous value.

3.5.1. Procedure

While there were no other cues to voicing for which measurements were available, previous research suggested that F1 onset frequency shows a trading relation with VOT similar in size to the one between VOT and VL (e.g., Summerfield & Haggard, 1977). Thus, the distributions used for this third cue were based on a small sample of acoustic measurements of F1 onset frequency for bilabial stops.[6] Mean values and standard deviations for the third cue are given in Table 1.

Training and testing procedures were the same as Simulation 2, except that the model had three cue-level MOGs. Learning rates were the same and initial parameters were determined in the same way as the first two simulations. After training, the models were tested on the VOT/VL pairs used in Simulation 2 under two conditions: (1) the artificial cue was held constant at an ambiguous value of 280 (constant-cue condition) or (2) the artificial cue covaried with VOT in nine steps from 240 to 320 (variable-cue condition).

3.5.2. Results and discussion

As in the previous simulations, the model performed quite well. Of the fifty models trained, only two were excluded because they overgeneralized. An additional two were excluded because they did not have any above-threshold categories in the combined MOG. Fig. 10 shows performance in the categorization task. In the constant-cue condition (panel B) a moderate VL effect is observed (similar to Simulation 2). In this condition, only VOT and VL are informative, and we see the predicted trading relation between the two cues. In the variable-cue condition (panel A), a decreased trading relation is observed, consistent with results from human listeners (J. C. Toscano & B. McMurray, un published data). This reflects the fact that the artificial cue is informative about the voicing category, decreasing the apparent effect of VL.

These results demonstrate that the size of trading relations can be changed without changes in cue weights. Because additional cues covaried with VOT, variation in responses along the VOT dimension reflected more than the contribution of VOT to the voicing judgment. VL, on the contrary, is uncorrelated with both of the other cues. As a result, the apparent size of the trading relation decreased. Cues are not weighted differ-

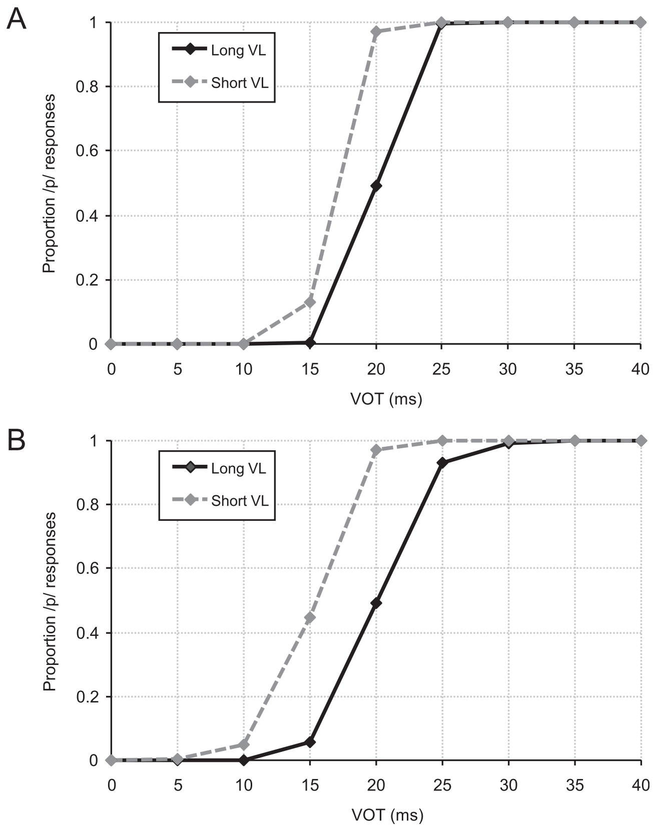

Fig. 10. Identification responses for Simulation 4. The trading relation is smaller when the third cue covaries with VOT (panel A) than when it is held constant (panel B).

ently, but, because variation in the primary dimension (VOT) now reflects variation in two cues, the overall contribution of that set of cues to the voicing judgment is greater than the contribution from VL alone. This results in a smaller trading relation between VOT and VL. Thus, both cue weights and the values of the cues used in testing determine the size of a trading relation.

4. General discussion

These simulations demonstrate that weighting-by-reliability, when adapted to the particular features of acoustic cues in speech, can be used to describe trading relations observed with human listeners. They also suggest that cue weights can be learned using a simple unsupervised competitive learning mechanism and that the learning process itself may play a role in determining how cues are weighted. Further, these models are not speech specific and could be applied to other categorization tasks as well.

The initial set of simulations revealed that the cue-weighting model provided a better fit to the data from listeners than the multidimensional model. The reason for this was counterintuitive: The cue-weighting model overweighted the less reliable cue, resulting in a trading relation. This was because of the fact that the categories and cue distributions are learned, not estimated directly from the input. Indeed, learning appears to be essential for obtaining the correct cue weights. Thus, the weights that listeners assign to cues may be a function of both the statistics of the input and the history of the learning system. While this result implies that the cue-weighting model represents certain cues suboptimally, this may generally be the best representation it can achieve given the requirements of learning. There may also be a benefit to this, in that it could allow the system to amplify cues that are generally weak and may be useful in other circumstances (e.g., a noisy environment in which VOT is hard to detect; Miller & Wayland, 1993).

While these results suggest that learning plays a role, we do not argue that any learning process will lead to this outcome. Other types of learning with different dynamics (Elman, 1993; McMurray et al., 2009b; Rumelhart & Zipser, 1985) might lead to different outcomes. Whether it comes from learning, or some other process, however, our simulations suggest that something must exaggerate the difference between overlapping distributions of VL. Statistics alone are not sufficient. Given the success of the MOG framework in accounting for a range of processes in speech development (de Boer & Kuhl, 2003; McMurray et al., 2009a; Vallabha et al., 2007), this approach provides a compelling explanation for listeners' performance in this task.

The final set of simulations revealed that changes in the trading relation between two cues can be observed without changes in the weights of the individual cues. Trading relations not only reflect cue weights but also the influence of correlated inputs. Preliminary work with human listeners (J. C. Toscano & B. McMurray, unpublished data) confirms the prediction of the cue-weighting model that changes in the VOT/VL trading relation can be observed in a single experiment when additional cues either covary with VOT or are held constant.

In addition to its close correspondence to behavioral data, the cue-weighting model provides a much more compact representation of the input than the multidimensional model because it collapses cues into a single dimension. This may offer a better approach for scaling up to a large number of cues. Further, this approach offers a general model of the origin of trading relations in speech, suggesting that they can largely be determined by the statistics of the input and unsupervised learning. This may allow us to explain trading relations between other sets of cues (Repp, 1982) as well as changes in cue weights over development (Mayo & Turk, 2004; Nittrouer, 2002).

The cue-weighting approach may also be informative for describing more general aspects of speech development. For example, during development, listeners face the problem of determining which cues are relevant for different phonological distinctions (Rost & McMurray, 2009). The cue-weighting model would suggest that the irrelevant cues simply receive a weight of zero (i.e., the model learns that they are best described as a single category). Thus, rather than first determining which cues are relevant and then learning the distribution of categories along those dimensions, learning may proceed by first determining the distributional statistics of a set of cues and then weighting them to determine if they are relevant (or both processes may happen simultaneously). This is a rather counterintuitive prediction, and it may be informative for understanding how listeners determine which cues to use to distinguish different phonological contrasts. It may also explain why infants at 14 months old who have tuned their speech categories to their native language, have still not entirely completed this process (J. C. Rost & B. McMurray, 2009, unpublished data).

4. 1. Relationship to other models

The cue-weighting model differs from previous approaches in several important ways. Both FLMP and NAPP models assume that cue integration is a category-dependent process. In contrast, our model suggests that cue weights can be determined independently of the categories along the dimension through an unsupervised learning process that uses the same information for learning the categories themselves. Although this contrasts with previous models, it provides a more realistic characterization of how cue weights are learned. In addition, there is evidence that integration may occur at precategorical stages, although it is not clear if this is because of the fact that cues are estimated and integrated, or whether these cues are only estimated in combination (Delgutte, 1982; Kingston & Diehl, 1994).

Recently, researchers have begun to use Bayesian (i.e., ideal observer) models (Griffiths & Tenenbaum, 2006; Tenenbaum & Griffiths, 2001) to describe various behaviors, including speech perception (Clayards et al., 2008; N. H Feldman, T. L. Griffiths, & J. L. Morgan, unpublished data; Norris & McQueen, 2008). These models share a number of properties with ours. They suggest that perceivers are sensitive to the distributional statistics of stimuli and use this information to categorize them. In addition, in both approaches, speech categories are described parametrically, allowing us to specify the properties of the model using a simple set of equations (Gaussian distributions in our case).

However, there are many aspects of our models that make them distinctly non-Bayesian. Bayesian models suggest that behavior is based on an optimal encoding of the statistics, whereas our simulations with the cue-weighting model and its close match to the behavioral data suggest that there are limits to how optimally these statistics are estimated. While perception is largely based on the distributional statistics of speech, we highlight a potentially important role for an iterative competitive learning process that leads to nonveridical perception of the input. This learning process eliminates the need to set the number of categories beforehand, using priors on *K* or pruning techniques. This emphasis on developmental plausibility has a further benefit: McMurray et al. (2009a) show how this iterative learning process can model the developmental time course of speech discrimination in infancy.

Other aspects, such as the decision rule, make our models suboptimal (Nearey & Hogan, 1986) and thus non-Bayesian. In addition, other features of Bayesian models, such as the *size principle* (i.e., learners decrease the size of a category with increasing exposure; Tenenbaum, 1999; Tenenbaum & Griffiths, 2001), are not required in our models and they may not obey certain constraints (the size of the category increases over training; see McMurray et al., 2009b). Again, this difference may arise from our commitment to iterative, developmental processes, but the fact that this model can learn quite successfully with rules that seem to violate this principle challenge whether it is necessary.

4.2. Limitations of the model

4.2.1. Perceptual and lexical processes

The MOG approach provides a transparent description of the structure of speech categories and explains how they can be derived from the statistics of the input using an unsupervised learning mechanism and simple form of competition. While this offers a good computational explanation of the system and a model of the *learning* process, it should not be taken as a model of the *perceptual* process, which may have additional effects on cue integration. For example, many cues (such as VOT and VL) are temporally asynchronous, and recent eye-tracking data suggest that listeners use cues as they become available during spoken word recognition, rather than waiting until all cues are received (McMurray et al., 2008). This suggests that the order in which cues are heard may have an effect on their functional weighting during perception. Indeed, the simulations presented here demonstrate that the developmental process can provide valuable insights about cue integration; an investigation of perceptual processing may yield further information. Thus, while the MOG provides a mechanistic account of learning, it provides only a descriptive account of listeners' perceptual processes. More detailed models of online speech processing may be needed to go further (see McMurray et al., 2009b; Toscano & McMurray, 2008).

A second aspect of speech processing that is not considered in these models is the role of feedback and top–down information in learning. The models presented here learn clusters of speech sounds based solely on bottom–up input. Lexical structure may be an important source of information for distinguishing speech sounds. For example, the fact that *bear* and *pear* are contrastive words in the lexicon may force the system to make fine-grained phonetic distinctions (Charles-Luce & Luce, 1990, 1995; Metsala & Walley, 1998; Walley, Metsala, & Garlock, 2003), and knowing which of the two words is being referred to can provide an error signal for supervised speech category learning (Kraljic & Samuel, 2006; Norris, McQueen, & Cutler, 2003). A complete model of speech development should include them. However, feedback is not necessary to account for the effects modeled here. Indeed, if lexical information was used to tag input with the correct category, we might not expect it to exaggerate the differences between the categories along the VL dimension as the cue-weighting model did and as human listeners do. Lexical information would help the model learn the *correct* distributions, decreasing the relative weight of VL. Thus, if this model included feedback its behavior would be less similar to listeners' behavior. This is, in effect, the result observed when the parameters of the Gaussians were set to match the distributional statistics of the input (discussed in Simulation 2). However, this does not rule out the utility of feedback. It may be necessary for learning other phonological distinctions, in particular, those for which there are no good individual cues.

4.2.2. Computational limitations

One limitation of the cue-weighing model is its ability to model sets of cues for which the relative order of categories along a dimension is different for each cue. For the case of VOT and VL, this problem can be solved by tracking the sign of the correlation between the two cues. However, for distinctions with more than two categories along each dimension (e.g., place of articulation in English, voicing in Thai), this solution will not necessarily work. For example, both VOT and F2 onset frequency are cues to word-initial place distinctions. However, the relative order of categories along the two dimensions is not the same in the context of the vowels /u/ and /o/ (Kewley-Port, 1982). The cue-weighting model would not be able to collapse these cues into a single dimension (although the weights of each dimension may still be accurate).

A related limitation is that the model cannot learn sets of cues in which the *within-category* correlations between cues (i.e., the correlation between tokens within each category) differ for each category. Fig. 11 shows a schematic representation of two hypothetical categories with different within-category correlations. Some pairs of cues show these types of relationships in different contexts (e.g., place of articulation: burst center frequency and F2 onset for different vowel contexts; Kiefte & Kluender, 2005; Nearey, 1998).

Both of these problems are a result of collapsing categories into a single dimension. The multidimensional model would not have these problems, since it can represent the entire acoustic space and categories may occur in any relative order along the cue dimensions. Thus, one solution would be to combine aspects of both models: reducing the number of cue dimensions as much as possible using the cue-weighting strategy outlined here and representing cues in a smaller multidimensional space. Another possible solution would be to take contextual information into account in determining the values along cue dimensions. That is, inputs to the cue dimensions could be determined by first partialing out context effects (see Cole, Linebaugh, Munson, & McMurray, in press).

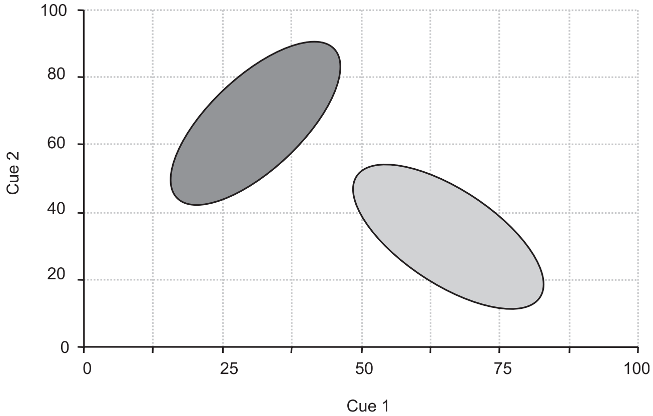

Fig. 11. A hypothetical pair of cues for which the within-category correlations for the two categories are different.

4.3. Conclusion

These simulations demonstrate the power of the weighting-by-reliability approach applied to speech. Weights given to acoustic cues can largely be determined from the statistics of the input, and these in turn do a good job predicting behavior. However, by themselves, these statistics are not sufficient to explain all types of cue weighting—some mechanism by which overlapping categories are enhanced is also needed. We suggest that learning itself, in addition to distributional statistics, may be crucial for determining those weights. While statistical learning approaches to perception have largely focused on *statistics*, there may also be a unique contribution for *learning*.

Notes

[1] All of the simulations were implemented in MATLAB. Code is available from the first author.

[2] For the single-cue case, we have explored a range of different starting μ values and learning rates (Toscano & McMurray, 2005). For initial μ values, we have not found major differences in the model's ability to learn or its categorization performance. Learning rates that are ≤1 tend to be the most successful.

[3] We used similar test stimuli to the ones used in McMurray et al. (2008), although they used different VLs for each of their word continua (because of variations in other phonological features of the words). The VL differences we chose span a similar range to the ones they used (100 ms in our simulations; 95–100 ms for McMurray et al.) and have similar values.

[4] We chose this loose definition of success because we wanted to include as many repetitions as possible to see how accurately the model reflected listeners' behavior. At minimum, the model needed to have at least two categories to compute identification functions for it. In addition, previous work has demonstrated that for reasonably distinguishable dimensions (e.g., VOT), MOG models show high accuracy in finding the correct parameters. For example, in models learning VOT, the average deviation from the correct VOT was 0.52 ms for μ (see McMurray et al., 2009a). For cases in which the model had more than two above-threshold categories at the end of training, the categories with the maximum posterior for the prototypical values of each cue (i.e., the mean value for each category) were used as the voiced and voiceless categories.

[5] Since the model normally starts with a large number of categories whose μ values are randomly distributed along the cue dimension, it is likely that some of these categories will fall outside this range, allowing the model to successfully learn the number of categories.

[6] While these values were not taken from a complete set of acoustic measurements for a third cue to voicing, we can examine the effects of an additional cue simply by allowing it to covary with VOT.

Acknowledgments

Portions of this work were presented at the 30th Meeting of the Cognitive Science Society, July 2008 in Washington, D. C. We are grateful to J. Sean Allen and Joanne Miller for acoustic data used to train the models. We would also like to thank Richard Aslin, Meghan Clayards, and Naomi Feldman for helpful theoretical discussions of the questions addressed in this paper, as well as Lisa Pearl and Terry Nearey for their comments on earlier versions of this manuscript. This research was supported by a University of Iowa Student Government Research Grant to J. C. Toscano and NIH DC008089 to B. McMurray.

References

Allen, J. S., & Miller, J. L. (1999). Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *Journal of the Acoustical Society of America*, *106*, 2031–2039.

Atkins, J. E., Fiser, J., & Jacobs, R. A. (2001). Experience-dependent visual cue integration based on consistencies between visual and haptic percepts. *Vision Research*, *41*, 449–461.

Battaglia, P. W., Jacobs, R. A., & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America*, *20*, 1391–1397.

de Boer, B., & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustic Research Letters Online*, *4*, 129–134.

Bülthoff, H. H., & Mallot, H. A. (1988). Integration of depth modules: Stereo and shading. *Journal of the Optical Society of America*, *5*, 1749–1758.

Charles-Luce, J., & Luce, P. A. (1990). Similarity neighbourhoods of words in young children's lexicons. *Journal of Child Language*, *17*, 205–215.

Charles-Luce, J., & Luce, P. A. (1995). An examination of similarity neighbourhoods in young children's receptive vocabularies. *Journal of Child Language*, *22*, 727–735.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*, 804–809.

Cole, J., Linebaugh, G., Munson, C., & McMurray, B. (in press). Vowel-to-vowel coarticulation across words in English: Acoustic evidence. *Journal of the Phonetics*.

Delgutte, B. (1982). Some correlates of phonetic distinctions at the level of the auditory nerve. In R. Carlson & B. Granstöm (Eds.), *The representation of speech in the peripheral auditory system* (pp. 131–150). Amsterdam: Elsevier Biomedical Press.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*, 71–79.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*, 429–433.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*, 752–782.

Frisch, S. (1996). *Similarity and frequency in phonology*. Unpublished doctoral dissertation, Northwestern University.

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

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*, 767–773.

Haggard, M. P., Ambler, S., & Callow, M. (1970). Pitch as a voicing cue. *Journal of the Acoustical Society of America*, *47*, 613–617.

Jacobs, R. A. (1999). Integration of texture and motion cues to depth. *Vision Research*, *39*, 3621–3629.

Jacobs, R. A. (2002). What determines visual cue reliability? *Trends in Cognitive Sciences*, *6*, 345–350.

Johnson, E. B., Cummings, B. G., & Landy, M. S. (1994). Integration of stereopsis and motion shape cues. *Vision Research*, *34*, 2259–2275.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America*, *108* (3), 1252–1263.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transaction of the ASME: Journal of Basic Engineering*, *82*, 35–45.

Kaufman, L. (1974). *Sight and mind*. New York: Oxford University Press.

Kewley-Port, D. (1982). Measurement of formant transitions in naturally produced stop consonant-vowel syllables. *Journal of the Acoustical Society of America*, *72*, 379–389.

Kiefte, M., & Kluender, K. R. (2005). Pattern playback revisited: Unvoiced stop consonant perception. *Journal of the Acoustical Society of America*, *118*, 2599–2606.

Kingston, J., & Diehl, R. (1994). Phonetic knowledge. *Language*, *70*, 419–454.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin and Review*, *13*, 262–268.

Landy, M. S., & Kojima, H. (2001). Ideal cue combination for localizing texture-defined edges. *Journal of the Optical Society of America*, *43*, 2307–2320.

Liberman, A. M., Harris, K. S., Kinney, J. A., & Lane, H. (1961). The discrimination of relative onset-time of the components of certain speech and nonspeech patterns. *Journal of Experimental Psychology*, *61*, 379–388.

Lisker, L. (1975). Is it VOT or a first-formant transition detector? *Journal of the Acoustical Society of America*, *57*, 1547–1551.

Lisker, L. (1986). ''Voicing'' in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech*, *29*, 3–11.

Lisker, L., & Abramson, A. S. (1964). A cross-linguistic of voicing in initial stops: Acoustical measurements. *Word*, *20*, 384–422.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2003). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309–332.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman.

Massaro, D. W., & Cohen, M. M. (1983). Categorical or continuous speech perception: A new test. *Speech Communication*, *2*, 15–35.

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

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generation. *Developmental Science*, *11*, 122–134.

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

Mayo, C., & Turk, A. (2004). Adult–child differences in acoustic cue weighting are influenced by segmental context: Children are not always perceptually biased toward transitions. *Journal of the Acoustical Society of America*, *115*, 3184–3194.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009a). Statistical learning of phonetic categories: Computational insights and limitations. *Developmental Science*, *12*, 369–378.

McMurray, B., Clayards, M., Tanenhaus, M. K., & Aslin, R. N. (2008). Tracking the timecourse of phonetic cue integration during spoken word recognition. *Psychonomic Bulletin and Review*, *15*, 1064–1071.

McMurray, B., Horst, J., Toscano, J. C., & Samuelson, L. K. (2009b). Towards an integration of connectionist learning and dynamical systems processing: Case studies in speech and lexical development. In J. Spencer,

M. Thomas, & J. McClelland (Eds.), *Toward a unified theory of development: Connectionism and dynamic systems theory re-considered* (pp. 218–249). London: Oxford University Press.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*, B33–B42.

Metsala, J. L., & Walley, A. C. (1998). Spoken vocabulary growth and the segmental restructuring of lexical representations: Precursors to phonemic awareness and early reading ability. In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 89–120). Mahwah, NJ: Erlbaum.

Miller, J. L., & Dexter, E. R. (1988). Effects of speaking rate and lexical status on phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 369–378.

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics*, *46*, 505–512.

Miller, J. L., & Wayland, S. C. (1993). Limits on the limitations of context-conditioned effects in the perception of [b] and [w]. *Perception & Psychophysics*, *54*, 205–210.

Nearey, T. M. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America*, *101*, 3241–3256.

Nearey, T. M. (1998). Locus equations and pattern recognition. Commentary on Sussman, Fruchter, Hilbert, & Sirosh, ‘‘Linear correlates in the speech signal: The orderly output constraint.’’ *Behavioral and Brain Sciences*, *21*, 241–259.

Nearey, T. M., & Assmann, 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. (1986). Phonological contrast in experimental phonetics: Relating distributions of measurements in production data to perceptual categorization curves. In J. Ohala & J. Jaeger (Eds.), *Experimental phonology*. New York: Academic Press.

Nittrouer, S. (2002). Learning to perceive speech: How fricative perception changes, and how it stays the same. *Journal of the Acoustical Society of America*, *112*, 711–719.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*, 357–395.

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

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

Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *Journal of the Acoustical Society of America*, *32*, 693–703.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency effects and the emergence of lexical structure* (pp. 137–157). Amsterdam: John Benjamins.

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

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, *82*, 81–110.

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*, 339–349.

Rumelhart, D. E., & Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science*, *9*, 75–112.

Schouten, B., Gerrits, E., & van Hessen, A. (2003). The end of categorical perception as we know it. *Speech Communication*, *41*, 71–80.

Shinn, P. C., Blumstein, S. E., & Jongman, A. (1985). Limitations of context conditioned effects in the perception of [b] and [w]. *Perception & Psychophysics*, *38*, 397–407.

Stevens, K. N., & Klatt, D. H. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. *Journal of the Acoustical Society of America*, *55*, 653–659.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *7*, 1074–1095.

Summerfield, Q., & Haggard, M. (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *Journal of the Acoustical Society of America*, *62*, 435–448.

Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*, *108*, 850–855.

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In M. S. Kearns, S. A. Solla, & D. A. Cohen (Eds.), *Advances in neural information processing systems 11*. Cambridge, MA: MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–640.

Toscano, J. C., & McMurray, B. (2005, November). *Statistical learning, cross-linguistic constraints, and the acquisition of speech categories: A computational approach*. Paper presented at the 11th Midcontinental Workshop on Phonology, University of Michigan, Ann Arbor, MI.

Toscano, J. C., & McMurray, B. (2008, November). *Online processing of acoustic cues in speech perception: Comparing statistical and neural network models*. Paper presented at the 156th Meeting of the Acoustical Society of America, Miami, FL.

Utman, J. A. (1998). Effects of local speaking rate context on the perception of voice-onset time in initial stop consonants. *Journal of the Acoustical Society of America*, *103*, 1640–1653.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 13273–13278.

Walley, A. C., Metsala, J. L., & Garlock, V. M. (2003). Spoken vocabulary growth: Its role in the development of phoneme awareness and early reading ability. *Reading and Writing*, *16*, 5–20.

Warren, P., & Marslen-Wilson, W. (1987). Continuous uptake of acoustic cues in spoken word recognition. *Perception & Psychophysics*, *41*, 262–275.

Wayland, S. C., Miller, J. L., & Volaitis, L. E. (1994). The influence of sentential speaking rate on the internal structure of phonetic categories. *Journal of the Acoustical Society of America*, *95*, 2694–2701.

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning & Development*, *1*, 197–234.

Werker, J. F., & Tees, R. C. (1984). Cross-langauge speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development*, *7*, 49–63.

Appendix

Bivariate (two-dimensional) Gaussian distribution for the multidimensional model

$$G_i(x,y) = \phi_i\left(\frac{1}{2\pi\sigma_{xi}\sigma_{yi}\sqrt{1-\rho_i^2}}\exp\left(-\frac{1}{2(1-\rho_i^2)}\left(\frac{(x-\mu_{xi})^2}{\sigma_{xi}^2}-\frac{2\rho_i xy}{\sigma_{xi}\sigma_{yi}}+\frac{(y-\mu_{yi})^2}{\sigma_{yi}^2}\right)\right)\right) \quad (A1)$$

This represents a single category in a two-dimensional mixture model, where μ_{xi} and μ_{yi} are the means along each dimension, σ_{xi} and σ_{yi} are the standard deviations along each dimension, ρ_i is the correlation between the two dimensions, and ϕ_i is the likelihood (frequency) of the category.

Learning rules for cue-weighting model

$$\Delta\phi_i = \eta_\phi \frac{G_i(x)}{M(x)} \tag{A2}$$

$$\Delta\mu_i = \eta_\mu \left(\frac{G_i(x)}{M(x)}\right) \frac{(x - \mu_i)}{\sigma_i^2} \tag{A3}$$

$$\Delta\sigma_i = \eta_\sigma \left(\frac{G_i(x)}{M(x)}\right) \left(\sigma_i^{-3}(x - \mu_i)^2 - \sigma_i^{-1}\right) \tag{A4}$$

These rules update the parameters of the Gaussians in the mixture such that they better approximate the distributions of the data on each training trial. In each equation, η is the learning rate for each parameter. *G*$_i$*(x)* is computed from Eq. 4 and *M(x)* is computed from Eq. 5. The update rule for each parameter is determined by taking the derivative of the likelihood function of the mixture distribution (Eq. 5) with respect to that parameter. Because *M(x)* is a sum, taking the partial derivative of any single parameter is only a function of the relevant category. This simplifies the learning rules, allowing us to drop all of the terms from the sum except for the one for the relevant category.

Learning rules for multidimensional model

$$\Delta\mu_{xi} = \eta_\mu \left(\frac{G_i(x)}{M(x)}\right) \frac{1}{(1 - \rho_i^2)} \left(\frac{x_j - \mu_{xi}}{\sigma_{xi}^2} - \frac{\rho_i y_j}{\sigma_{xi}\sigma_{yi}}\right) \tag{A6}$$

$$\Delta\sigma_{xi} = \eta_\sigma \left(\frac{G_i(x)}{M(x)}\right) \left(\frac{(x_j - \mu_{xi})^2}{\sigma_{xi}^3(1 - \rho_i^2)} - \frac{\rho(x_j - \mu_{xi})(y_j - \mu_{yi})}{\sigma_{xi}^2\sigma_{yi}(1 - \rho_i^2)} - \frac{1}{\sigma_{xi}}\right) \tag{A7}$$

$$\begin{aligned}\Delta\rho_i = \eta_\sigma \left(\frac{G_i(x)}{M(x)}\right) \left(\frac{1}{1 - \rho_i^2}\right) \Bigg(\rho_i^3 - \left(\frac{1}{1 - \rho_i^2}\right) \Bigg(&\frac{\rho_i(x_j - \mu_{xi})^2}{\sigma_{xi}^2} \\ &- \frac{(\rho_i^2 + 1)(x_j - \mu_{xi})(y_j - \mu_{yi})}{\sigma_{xi}\sigma_{yi}} + \frac{\rho_i(y_j - \mu_{yj})^2}{\sigma_{yi}^2}\Bigg)\Bigg)\end{aligned} \tag{A8}$$

As with the learning rules for the cue-weighting model, η is the learning rate for each parameter, and the update rules are computed by taking the derivative of the likelihood function of the mixture distribution (Eq. A1) with respect to each parameter.