

Using the Distributional Statistics of Speech Sounds for Weighting and Integrating Acoustic Cues

Joseph C. Toscano (joseph-toscano@uiowa.edu)
University of Iowa, Department of Psychology, E11 SSH
Iowa City, IA 52242 USA

Bob McMurray (bob-mcmurray@uiowa.edu)
University of Iowa, Department of Psychology and Iowa Center for Developmental and Learning Sciences, E11 SSH
Iowa City, IA 52242 USA

Abstract

A great deal of behavioral evidence suggests that infants can use distributional statistics to learn speech sound categories. Recently, a number of computational approaches have demonstrated the feasibility of statistical learning by showing that the distributional statistics of linguistically-relevant acoustic cues can be learned in an unsupervised way. However, speakers and listeners use a large number of acoustic cues to distinguish phonetic categories, and it is not clear how multiple cues are combined during perception. We propose a model of speech sound category acquisition that learns the distributions of multiple cues that lie along the same dimension and combines them. We demonstrate that the model is able to account for trading relations between cues (an indicator of the size of the effect of each cue) for word-initial voicing contrasts in English.

Keywords: speech perception; speech development; mixture of Gaussians; cue integration; statistical learning.

Introduction

The sound systems of human languages vary greatly. One of the first steps in language acquisition is for infants to determine the sound structure of their native language. In a given language, phonetically-relevant acoustic cues are distributed such that they tend to cluster into categories. For example, voice-onset time (VOT; the delay between the opening of the vocal tract and the onset of vocal energy) values in English tend to cluster into voiced and voiceless categories, near 0 and 50 ms, respectively (Lisker & Abramson, 1964). Thus, the distributional statistics of this cue contains information about the voicing categories of English.

Previous work has demonstrated that infants can track these distributional statistics and use this information to learn categories (Maye, Werker, & Gerken, 2002). Recently, researchers have begun to use computational models to understand this process more deeply. These models have been used to describe how learning unfolds over development and leads to stable speech sound categories.

One way to model this process is to represent each phonetic category (such as *voiced* or *voiceless*) as a Gaussian distribution, providing us with a representation that corresponds to the frequency distribution of an acoustic cue. McMurray, Aslin, and Toscano (in press) present a model that uses this approach. Figure 1A shows how the model might represent the VOT distribution of English. It contains two categories, one centered at the mean of the voiced VOT values and the other centered at the mean of the voiceless values. The model

learns the number of categories in the input and their statistical distributions, reflecting the developmental trajectory of speech category formation in the process. This demonstrates that unsupervised statistical learning mechanisms are able to describe how infants can acquire these categories.

However, many critical aspects of this process have yet to be addressed. In particular, acoustic analyses have revealed that most phonemic distinctions are marked by multiple cues. Lisker (1978) cites that there are at least 16 acoustic cues that distinguish voiced and voiceless sounds in word-medial position in English. Existing models of statistical learning and speech sound category acquisition are insufficient to describe how listeners learn multiple cue distributions and integrate these cues when perceiving speech. These challenges are also not limited to understanding speech development and perception. Similar difficulties would be encountered when trying to combine multiple features in other domains as well.

Can statistical learning approaches be extended to capture listeners' use of multiple acoustic cues? How could this problem be instantiated in a model of speech sound categorization? One way would be to present each cue along a separate dimension, as in Figure 1B, leading to n -dimensional categories, where n is the number of cues to be learned. This type of model has the advantage of allowing us to fully represent the acoustic space. However, it can also lead to computational complexity (representing a 16-dimensional category for the cues to voicing), and it would present the learner with a sparse space from which to extract the categories, since many of the possible combinations of cues would never be heard.

An alternative approach would be to weight and combine sets of cues that lie along the same phonetic dimension, as shown in Figure 1C. Cues to voicing, for example, are not orthogonal; they each indicate the same category structure: voiced or voiceless for English. Indeed, they are similar to visual cues to depth that observers integrate during perception to determine the three-dimensional structure of objects (Jacobs, 2002). Thus, it is possible to integrate them into a single phonological dimension upon which the relevant categories can be learned. This approach reduces the computational complexity of the problem, and it can be expanded to account for a large number of acoustic cues.

This paper seeks to address several questions about how sets of multiple cues can be learned:

- Can we base a statistical learning model of speech sound categories on a weighted combination of cues or will only a full (multi-dimensional) model work?
- Can such a model account for the available data?
- What does this tell us about cue integration principles?

We present two models and simulations designed to answer these questions. The models are mixtures of Gaussians (MOGs) that have been used previously to model the acquisition of a single acoustic cue (McMurray, Aslin, & Toscano, in press) and multiple cues along orthogonal dimensions (Valabha, McClelland, Pons, Werker, & Amano, 2007). Here, we will examine how these types of models can be used to learn a set of cues along a single sub-phonemic dimension.

Models

We contrast two models of speech sound categorization that attempt to demonstrate how listeners can learn and combine

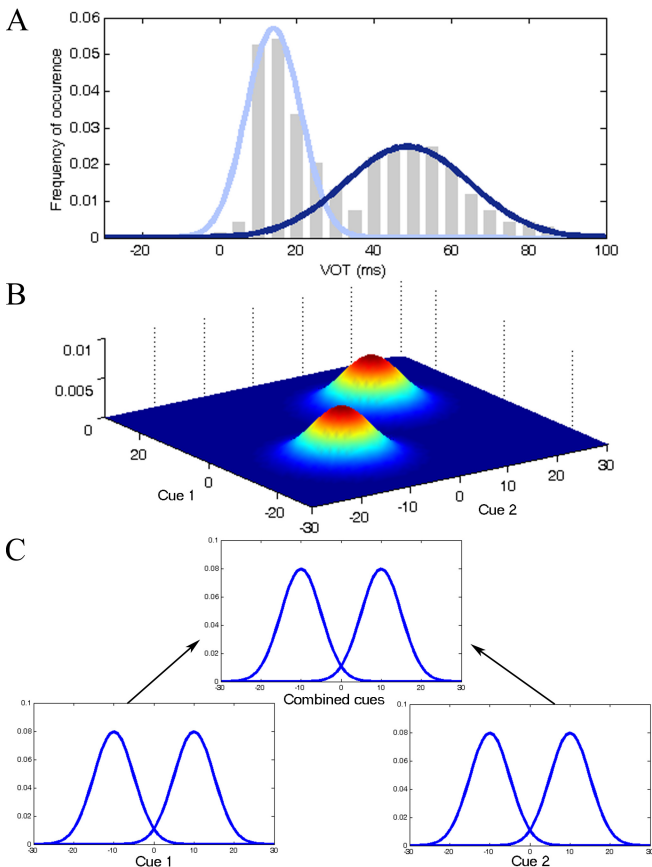


Figure 1: (A) VOT distribution of English represented by two Gaussian distributions corresponding to voiced (lighter color) and voiceless (darker) categories. The gray bars show the likelihood of VOT values obtained from the acoustic measurements in Allen and Miller (1999). (B) A representation of categories along two cue dimensions. (C) Integration of multiple cues into a single dimension.

multiple acoustic cues. The two models differ in how they combine information from multiple cues. The *cue weighting model* weights cues and sums the weighted estimates provided by each cue, reducing them to a single dimension (whose distribution is learned by another MOG). The *multi-dimensional model* instead uses a set of two-dimensional Gaussians to track the combined distribution of cue-values, representing all possible combinations of cues.

Cue weighting model

Architecture The model consists of several mixtures of Gaussian distributions. Each MOG contains a series of K Gaussians along a particular acoustic dimension. Each Gaussian represents a potential phonetic category. Since the number of categories is not known beforehand and must be learned over development, the mixture contains more Gaussians than it needs. One problem with MOG models is determining the correct number of Gaussians. To solve this problem, each Gaussian contains a frequency parameter, ϕ , corresponding to its prior probability. The model can then reduce the ϕ -values of categories that are not needed.

The likelihood of a particular value along that cue dimension, for each Gaussian (i) is defined by the posterior of that Gaussian times its ϕ -value:

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

where μ is the mean of the distribution and σ is the standard deviation. The sum of the probabilities for each Gaussian in the mixture determines the likelihood of a cue value:

$$M(x) = \sum_i^K G_i(x) \quad (2)$$

For example, if a MOG represents the voicing categories for English along the VOT dimension, it would contain two Gaussians ($K=2$) – one corresponding to the voiced category ($\mu=0, \sigma=10$) and one corresponding to the voiceless category ($\mu=50, \sigma=20$) with equal prior probabilities. The likelihood of a 0 ms VOT would be the sum of the relatively high probability for the voiced category and the relatively low probability for the voiceless category. The likelihood of a 20 ms VOT, in contrast, would be the sum of the low probabilities from both categories. Thus, the category structure of this model would match the structure of the VOT categories for English.

The model contains a MOG for each acoustic cue and an additional MOG for representing the categories based on the combination of the cues for the phonetic distinction being learned.

Learning The model learns the category structure of an acoustic cue by adjusting the parameters of the Gaussians in the mixture for that cue. Learning is accomplished via maximum likelihood estimation by stochastic gradient descent.

As mentioned above, the mixture represents the likelihood of a particular cue value given a set of parameters (μ, σ , and

ϕ). Since learning is iterative and we assume there are no priors on the parameters, Bayes' theorem says that the mixture also represents the likelihood of a set of parameters given a particular cue value. We can therefore use gradient descent to update the parameters of each Gaussian given a particular data point. Gradient descent updates the Gaussians by the derivative of the likelihood function, (2), with respect to each parameter (See McMurray, Horst, Toscano, and Samuelson (in press) for the learning rules used in the model).

Learning proceeds by presenting the model with individual cue values, calculating the change in each parameter value, and updating the parameters. The ϕ values are normalized so they sum to one and reflect the likelihood of each category.

The model uses winner-take-all competition so that ϕ is only changed for the Gaussian that has the highest likelihood. This solves the problem mentioned above of not knowing the number of categories *a priori*. Competition is necessary for the model to determine the correct number of categories along each dimension (McMurray, Aslin, & Toscano, in press). In addition, this is psychologically plausible, since a particular input corresponds to only a single category, and learning only needs to occur for that category.

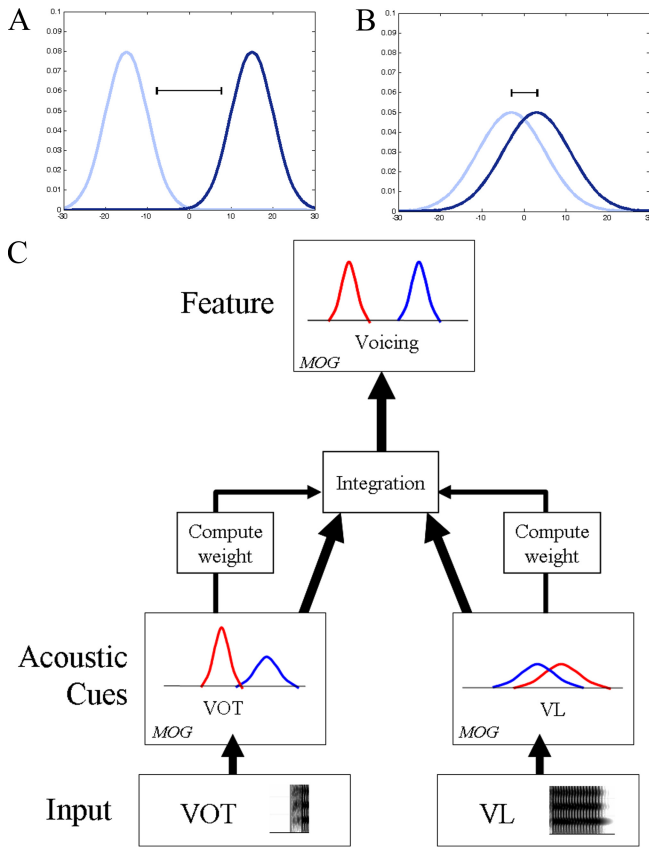


Figure 2: (A) More reliable categories (solid lines) with a high overall σ . The line in the center represents the approximate overall σ . (B) Less reliable categories with a lower overall σ . (C) Schematic diagram of the cue weighting model.

Cue weighting and integration In order to implement cue integration in the model, this learning procedure is applied to multiple MOGs representing different acoustic cues for a given phonetic contrast. Cue integration occurs by weighing the input along each dimension and summing the weighted cue values, which serve as input to a separate MOG that lies along a dimension corresponding to the phonetic distinction being learned. This MOG is also trained and produces a set of speech categories that contain information from multiple cues. The individual MOGs for each cue are only used to compute the weights and inputs to the combined MOG.

As with the parameters of the Gaussians, cue weights are computed on the basis of the distributional statistics of the input. Cues with a higher reliability receive a greater weight. This approach is similar to a Kalman filter (Kalman, 1960; Jacobs, 2002), which computes reliability for the individual cues and then uses a linear combination rule to integrate data from each source:

$$x = \sum_i^n w_i x_i \quad (3)$$

where x is the estimate based on the combined input, x_i is the estimate for a particular cue (i), w_i is the weight for that cue, and n is the number of cues. In a Kalman filter, weights are determined by the formula

$$w_i = \frac{1}{\sigma_i^2} \bigg/ \sum_j^n \frac{1}{\sigma_j^2} \quad (4)$$

where σ^2 is the variance of the distribution along a particular cue dimension. Thus, cues that are more reliable (i.e. those with a smaller variance) will have larger weights.

This measure of reliability has been used previously to model sensory integration given unimodal cue distributions along a common dimension (Jacobs, 2002). For the types of distributions we are examining here, however, the overall variance cannot be used, since each cue contains multiple Gaussians. A single variance estimate does not adequately describe a dimension that is distributed in this way, since a high variance could be achieved by a mixture of two narrow Gaussians in which the means are far apart (highly reliable, Figure 2A) and a low variance could be achieved by a mixture of two broad Gaussians that are highly overlapping (unreliable, Figure 2B). Thus, we developed a new estimate of cue reliability that can be used to determine the weight of a multimodal distribution:

$$g = \left(\sum_n^K \sum_m^K \frac{\phi_m \phi_n (\mu_m - \mu_n)^2}{\sigma_m \sigma_n} \right) / 2 \quad (5)$$

This equation takes into account the variance of each distribution along a dimension, as well as the means and prior probabilities for each distribution. It calculates the reliability of a particular cue dimension by summing all pairwise comparisons between the Gaussians along that dimension (m and

n refer to particular Gaussians; K refers to the total number of Gaussians along that cue dimension). Pairs of Gaussians that are close together do not contribute much to g , since the difference between their μ s is minimal. Similarly, pairs of Gaussians with high σ values do not contribute as much as pairs with low σ s, and low frequency (ϕ) Gaussians do not contribute much to g . This results in a measure of overall reliability of a dimension analogous to d' . (6) normalizes the reliability estimates to compute the weight for each cue:

$$w_i = \frac{g_i}{\sum_j g_j} \quad (6)$$

Simulation procedure For a single run of the model, training data is generated by randomly sampling from distributions corresponding to the speech categories of interest. The model is initialized by randomly setting μ for K Gaussians to a value in the range of the data to be learned. σ is set to a constant value for each Gaussian, and ϕ is set to $1/K$. On each trial, the model is presented with individual exemplars to each MOG representing an acoustic cue, and the parameters of each Gaussian are updated using the learning rules and competition described above. Next, the cue dimensions are weighted, and the input along each dimension is normalized relative to the parameters of the Gaussians along that dimension. A new input is computed using the input to the individual cues and their weights. The MOG representing the combined percept receives this as input, and learning occurs along this dimension. Figure 2C shows the basic organization of a model that integrates two cues to voicing, VOT and vowel length (VL).

After training, the model is tested on its categorization of stimuli varying along each acoustic dimension. From this, we can see how it uses multiple cues by measuring trading relations, differences in categorization depending on the values of the cues.

Multi-dimensional model

Architecture The multi-dimensional model uses the same basic MOG framework as the cue weighting model. However, rather than representing each cue along a separate dimension and combining them, this model consists of a mixture of n -dimensional Gaussians. For the simulations presented here, a two-dimensional MOG is used. The likelihood of a particular set of cue values for each Gaussian is defined by

$$G_i(x) = \phi_i \left(\frac{1}{2\pi|\Sigma_i|^{1/2}} \exp \left(-\frac{1}{2} (x - \mu_i)^\top \Sigma_i^{-1} (x - \mu_i) \right) \right) \quad (7)$$

where Σ is the covariance matrix for the two cues. Other parameters are the same as those in (1) for each cue. As in the cue weighting model, the overall likelihood of a set of cue values is defined by (2). Figure 1B shows how this model might represent categories defined by two cues.

Learning and simulation procedure Learning and testing follow the same basic procedure as in the cue weighting model. On each trial, a pair of cue values is given as input and the parameters of the MOG are updated.

Simulations

Simulation 1: Two cues

The first simulation involved a trading relation between two acoustic cues and was designed to determine whether the cue weighting or multi-dimensional model better accounts for cue integration in speech along a single phonetic dimension.

The cue weighting model contained three MOGs: one for representing VOT, one for vowel length (VL), and one for representing overall voicing based on both cues. The multi-dimensional model contained a single two-dimensional MOG, with one dimension for VOT and one for VL. 50 models of each type were run in the simulation.

Training The models were trained on VOT and VL values randomly sampled from distributions based on the acoustic measurements from Allen and Miller (1999) for VL and Lisker and Abramson (1964) for VOT. The models were run for a sufficient amount of time for them to settle on a stable set of categories. A total of 70,000 trials were run for the cue weighting model, and 200,000 trials were run for the multi-dimensional model. For the correlation between the cues used by the multi-dimensional model, the value from the Allen & Miller dataset was used. Table 1 shows the means and standard deviations of the distributions used to generate the training data. Models that overgeneralized, that is, those that had only a single category after training for any of the MOGs (as determined by the number of Gaussians with ϕ -values greater than 0.1) were excluded from analysis.

Testing Each model was tested on the VOT and VL values used with human listeners in McMurray, Clayards, Tanenhaus, and Aslin (submitted): nine VOT steps (from 0 to 40 ms) and two VL values (125 and 225 ms). The /b/ and /p/ categories were identified by finding the Gaussians with the highest posterior probabilities for the best /b/ exemplar (VOT=0 ms; VL=225 ms) and the best /p/ exemplar (VOT=40 ms; VL=125 ms) in either the combined MOG for the cue weighting model or the two-dimensional MOG in the multi-dimensional model. These two Gaussians were used to compute the likelihood of a /p/ response for each stimulus with the

Table 1: Descriptive statistics of distributions used to generate training data. Means and standard deviations are in ms for VOT and VL and Hz for F1.

	Voiced			Voiceless		
	VOT	VL	F1	VOT	VL	F1
Mean	0	188	260	50	170	300
SD	5	45	10	10	44	10

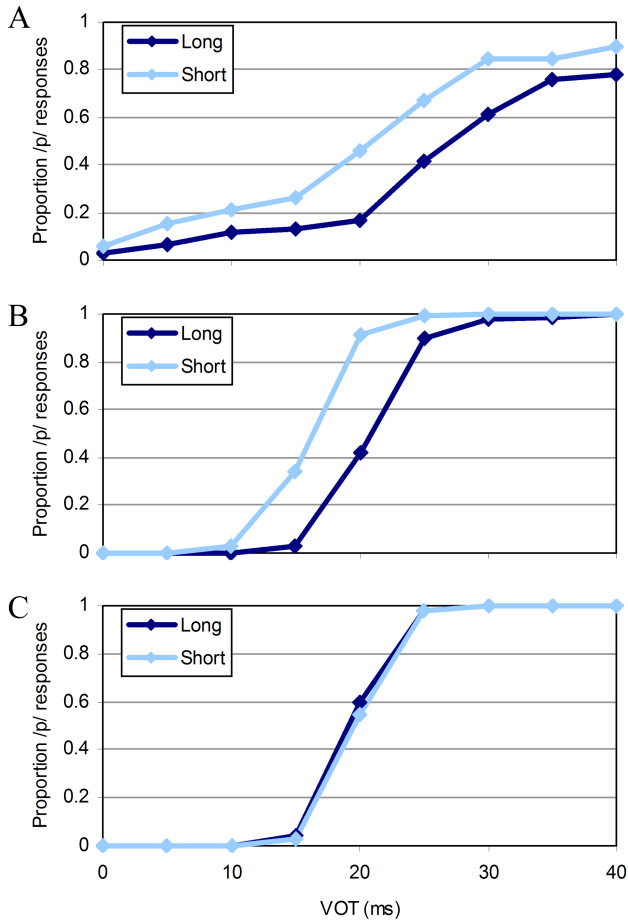


Figure 3: (A) Data from human listeners from a picture identification task with stimuli varying in VOT and VL. (B) Data from the cue weighting model. (C) Data from the multi-dimensional model.

Luce choice rule. This produced a response corresponding to the proportion of /p/ responses obtained for human listeners.

Results Figure 3A shows behavioral data from human listeners from McMurray et al. (submitted), displaying proportion of /p/ responses as a function of VOT and VL. A moderate shift of the VOT category boundary of about 5 ms for the two different VL conditions can be seen. Figures 3B and 3C show the results from the simulations with the cue weighting and multi-dimensional models, respectively. The results from the cue weighting model show a similar-sized shift in the VOT boundary to the human data, but the multi-dimensional model shows no observable shift in the predicted direction. The RMS difference between the human data and cue weighting models is 0.141, and between the human data and multi-dimensional models is 0.183.

The results of the simulations indicate that the cue weighting model provides a better fit to human listeners' categorization. The multi-dimensional model did not show the predicted VL effect, suggesting that it did not use this less discriminable cue and instead relied on VOT for categorization.¹

Simulation 2: Three cues

The second set of simulations was designed to determine if the cue weighting model could also account for changes in the size of trading relations depending on the value of additional acoustic cues in the signal. Toscano and McMurray (in preparation) found that the size of the trading relation between VOT and VL was dependent on whether a third cue to voicing, F1 at voicing onset, was ambiguous or informative. In natural speech, this cue covaries with VOT, reducing the apparent size of the VOT/VL trading relation. Toscano and McMurray (in preparation) examined this in human listeners using synthetic speech that contained formant onsets that either covaried with VOT (similar to natural speech) or were held constant at an ambiguous value. This simulation will look at whether the cue weighting model produces a corresponding change in the size of the trading relation.

50 models with four MOGs (VOT, VL, F1, and a combined MOG), were trained and their categorization was tested for F1 values that covaried with VOT or were held constant.

Training VOT, VL, and F1 values were randomly sampled from the distributions in Table 1. F1 values were estimated from acoustic measurements of the stimuli in Toscano and McMurray (in preparation). Each model was run for 90,000 trials. Other parameters were identical to Simulation 1.

Testing The testing procedure was the same as in Simulation 1, except that the model was presented with three acoustic cues. The model was tested under two conditions. In one condition (constant), F1 values were held constant at 280 Hz. In the second condition (covaried), F1 values covaried with VOT in 10 Hz increments from 240 to 320 Hz.

Results Figure 4A shows the results for human listeners' from Toscano and McMurray (in preparation) and Figure 4B shows the results from the cue weighting model. The model shows a reduced trading relation between VOT and VL when F1 covaries with VOT, replicating the basic effect observed with human listeners. These results indicate that the cue weighting model was able to account for the difference in the size of the trading relation observed in the human data.

General Discussion

The results of these simulations suggest that the cue weighting model and the reliability metric used here provide a reasonable account of listeners' performance. They also suggest that a full model of the entire acoustic space may not be necessary, and, in fact, may not correctly represent the weights that listeners assign to acoustic cues. Further, these simula-

¹In both models, Gaussians tended to not overlap significantly. This led to a negligible VL effect in the multi-dimensional model. Because Gaussians could be separated along both cue dimensions, the model could approximate the close means of the VL data without having overlapping categories. In contrast, VL means in the cue weighting model were further apart than the means in the data, since the model could only reduce overlap for those Gaussians along that dimension. This produced a larger VL effect than we would expect based on the cue weighting metric alone.

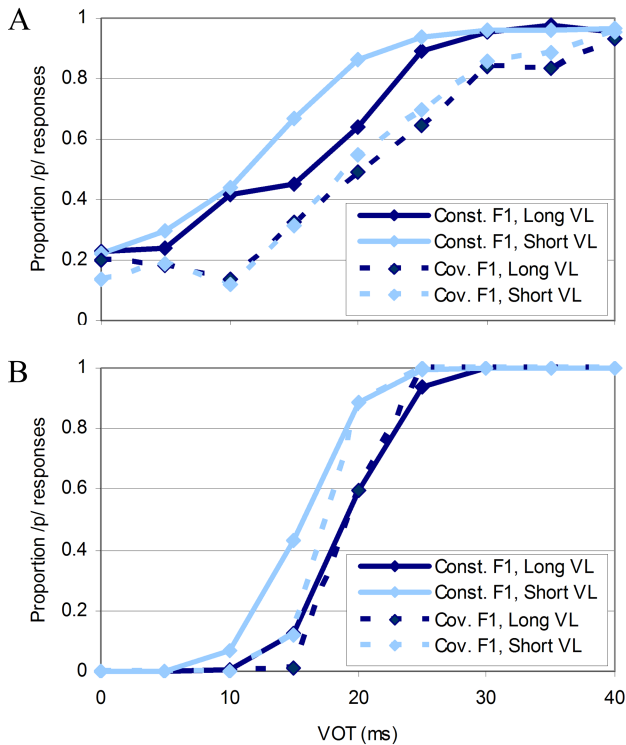


Figure 4: Identification responses for three cues to voicing for either covaried or constant F1 values. (A) Data for human listeners from Toscano and McMurray (in preparation). (B) Data from the cue weighting model from Simulation 2.

tions suggest that the information needed to weight cues is available in the statistics of the input – information that infants already use to discover the category structure of individual acoustic cues.

A multi-dimensional space may be used to map acoustic information onto speech categories at other levels of processing. Indeed, this approach seems particularly useful for combining cues across orthogonal dimensions, such as manner, place and voicing for consonants, or frontness and backness for vowels (see Vallabha et al. (2007) for simulations with a three-dimensional MOG learning vowel spaces). However, this would not be needed if listeners only had to learn phonetic features rather than phoneme-like units and could base perception on these features. Note also that the cue weighting model may not be able to successfully combine cues if the within-category correlations between the cues are highly different (since it does not track these correlations). However, we know of no set of acoustic cues for which this is the case.

The advantage of the weighting and integration approach lies in the fact that a large number of acoustic cues can be combined into a simpler representation. For a distinction such as word-medial voicing, in which there are at least 16 cues that contain the same phonetic category structure, it may be difficult to learn the distributions of categories in a 16-dimensional space. By combining cues into a single dimen-

sion, this reduces the computational difficulty of learning speech sound categories.

The cue weighting model presented here extends the statistical learning framework to explain how the distributions and weights of multiple cues can be learned over development. Also, the reliability metric used here provides a general method for weighting a multimodal distribution. This approach may be useful for understanding other types of feature combination as well.

Acknowledgments

We would like to thank J. Sean Allen and Joanne Miller for the acoustic data used to train the model. This research was supported by a University of Iowa Student Government Research Grant to JCT and NIH DC008089 to BM.

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.

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

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

Lisker, L. (1978). Rapid vs. rabad: A catalogue of acoustic features that may cue the distinction. *Haskins Laboratories Status Report on Speech Research*, SR-54, 127-132.

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

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

McMurray, B., Aslin, R. N., & Toscano, J. C. (in press). Statistical learning of phonetic categories: Computational insights and limitations. *Developmental Science*.

McMurray, B., Clayards, M., Tanenhaus, M., & Aslin, R. (submitted). Tracking the time course of phonetic cue integration during spoken word recognition.

McMurray, B., Horst, J., Toscano, J. C., & Samuelson, L. K. (in press). Towards an integration of connectionist learning and dynamical systems processing: case studies in speech and lexical development. In J. P. Spencer, M. Thomas, & J. McClelland (Eds.), *Toward a new grand theory of development: Connectionism and dynamic systems theory reconsidered*. Oxford University Press.

Toscano, J. C., & McMurray, B. (in preparation). Integration of temporally asynchronous cues to voicing in natural and synthetic speech.

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*, 104, 13273-13278.