SPECIAL SECTION: COMPUTATIONAL PRINCIPLES OF LANGUAGE ACQUISITION

Statistical learning of phonetic categories: insights from a computational approach

Bob McMurray,[1] Richard N. Aslin[2] and Joseph C. Toscano[1]

1. Department of Psychology and the Delta Center, University of Iowa, Iowa City, USA
2. Department of Brain and Cognitive Sciences, University of Rochester, Rochester, USA

Abstract

Recent evidence (Maye, Werker & Gerken, 2002) suggests that statistical learning may be an important mechanism for the acquisition of phonetic categories in the infant's native language. We examined the sufficiency of this hypothesis and its implications for development by implementing a statistical learning mechanism in a computational model based on a mixture of Gaussians (MOG) architecture. Statistical learning alone was found to be insufficient for phonetic category learning – an additional competition mechanism was required in order for the categories in the input to be successfully learnt. When competition was added to the MOG architecture, this class of models successfully accounted for developmental enhancement and loss of sensitivity to phonetic contrasts. Moreover, the MOG with competition model was used to explore a potentially important distributional property of early speech categories – sparseness – in which portions of the space between phonetic categories are unmapped. Sparseness was found in all successful models and quickly emerged during development even when the initial parameters favoured continuous representations with no gaps. The implications of these models for phonetic category learning in infants are discussed.

Introduction

Infants face a difficult problem in acquiring their native language because the acoustic/phonetic variability in the input far exceeds the limited number of distinctive differences that the phonemes of their language. How do infants attend to the relevant information that distinguishes words? Recent evidence suggests that phonemic categories may be induced, in whole or in part, by a rapid statistical learning mechanism that is sensitive to the distributional properties of phonetic input (Maye, Werker & Gerken, 2002; Maye, Weiss & Aslin, 2008). This evidence suggests that the detailed frequency-of-occurrence of tokens along continuous speech dimensions plays a crucial role in the formation and modification of phonemic categories.

The present paper describes a computational model of statistical speech category learning that examines the necessary and sufficient mechanisms needed to account for known empirical data from infants, and the implications of those mechanisms for early speech categories. We demonstrate that statistical learning alone is insufficient: competition is also required. However, once this feature is added to the model, it can account for a number of developmental trajectories in speech category learning. Finally,

we examine the possibility that early speech categories are independent and *sparsely* distributed; that is, they do not fully cover all values along a phonetic dimension.

Statistical learning and development

The classic view of speech perception in both adults (cf. Liberman, Harris, Hoffman & Griffith, 1957) and infants (cf. Eimas, Siqueland, Jusczyk & Vigorito, 1971; see Jusczyk, 1997) is that stop consonants are perceived categorically. However, more recent evidence confirms within-category sensitivity in both adults (Pisoni & Tash, 1974; Carney, Widen & Viemeister, 1979; Miller, 1997) and infants (Miller & Eimas, 1996; McMurray & Aslin, 2005). Nevertheless, adults and infants have a bias to group acoustically similar sounds into categories, and these categories begin to match native language by 6 to 12 months of age. This matching process can take a number of forms. For some dimensions, infants are initially able to distinguish a number of contrasts not found in their native language, and this is followed by a loss of the unnecessary contrasts (Werker & Tees, 1984; see Werker & Curtin, 2005 and Kuhl, 2004 for recent reviews). For others, contrasts are initially indiscriminable but are
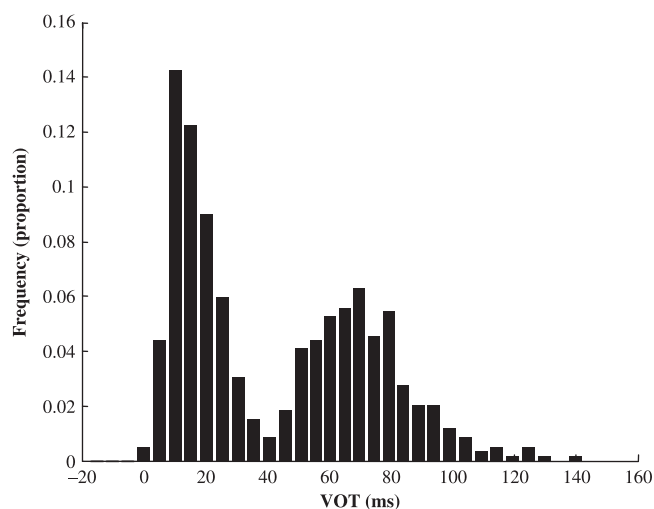
Figure 1 *The bimodal distribution of voice onset time (VOT) in English. Shown is the relative frequency as a function of VOT collapsed across all three places of articulation. Two clusters are clearly visible: one centred around 10–15 ms (Voiced) and one centred around 65–75 ms (Voiceless). Data are from Allen & Miller (1999).*

enhanced over development (Eilers & Minifie, 1975; Eilers, Wilson & Moore, 1977).

Speech exemplars tend to cluster statistically along continuous acoustic/phonetic dimensions. For example, Figure 1 shows measurements of voice onset times (VOTs: distinguishing voiced from voiceless sounds) from English speakers, illustrating two clusters corresponding to voiced and voiceless categories (Lisker & Abramson, 1964; Allen & Miller, 1999). These clusters approximate Gaussian distributions, one centred at 0 ms (voiced) and one at 60 ms (voiceless). Whereas the incidence of tokens near 0 and 60 ms is frequent, few are attested at 30, 100 and –20 ms. A similar clustering is seen for the cues to vowels (Peterson & Barney, 1951; Hillenbrand, Getty, Clark & Wheeler, 1995) and approximants (Espy-Wilson, 1992).

This suggests that the frequency of occurrence of tokens along a given speech dimension could allow listeners to induce phonetic categories from the clusters of tokens in the input. This has been explicitly tested behaviourally with adults (Maye & Gerken, 2000) and infants (Maye *et al.*, 2002, 2008). Listeners who were exposed to a series of speech sounds for which the frequency of any VOT was distributed bimodally (characteristic of two categories) were able to discriminate two exemplars that straddled the category boundary, whereas listeners exposed to a unimodal (single category) distribution could not. Thus, a few minutes of exposure to statistically structured input biases perception in a way that is consistent with statistical learning.

Existing models

A variety of computational models implement category learning via clustering algorithms. Connectionist models

(Elman & Zipser, 1986; Guenther & Gjaja, 1996; Nakisa & Plunkett, 1998; see also McCandliss, Fiez, Protopapas, Conway & McClelland, 2002) have demonstrated the feasibility of input-driven learning mechanisms, but such models incorporate other features that make it difficult to isolate statistical learning. Elman and Zipser (1986) used nonlinear activation functions and competition (dimensionality reduction); Guenther and Gjaja (1996) employed topographic competition; and Nakisa and Plunkett's (1998) genetic algorithm produced an array of specialized architectures and learning rules. In these models, statistical learning is one of many mechanisms involved in speech category development.

To better isolate the clustering mechanism, we simulated statistical learning in a model that uses a simple architecture, makes few theoretical assumptions, and adds constraints only when needed to account for the data. Many models start with a theoretical paradigm (e.g. connectionism or dynamical systems theory) and ask whether the principles of this paradigm are sufficient to solve a problem. Our goal was to start with the computational problem the system is trying to solve (learning the mapping between continuous inputs and categories) and use current theory (distributional learning) to arrive at an architecture suited to that problem.

This led us to use the mixture of Gaussians (MOG) approach. This is a classic tool from statistics and computer science used for estimating the parameters of a set of probabilistic clusters (Titterington, Smith & Makov, 1985). Although this requires certain architectural assumptions, these are made with respect to the problem being solved, not as a result of any paradigmatic approaches to development. This approach allows us to isolate and evaluate statistical learning as a mechanism for forming categories.

The mixture of Gaussians model

The MOG approach to speech categorization estimates the probability, *M*(*vot*), of obtaining any individual cue-value (e.g. a specific VOT) as the sum of the probability it came from some number (*K*) of overlapping Gaussians, each with some prior likelihood. The model represents each potential or actual category (e.g. voiced or voiceless) with a single Gaussian (G_i) that has a set of parameters that describe the frequency of that category (ϕ), its location in the input-space (μ), and its variability in the input-space (σ) (see Figure 2, Equation 1 for an example along the VOT dimension):

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

Each Gaussian computes the likelihood of hearing a specific VOT if that category was the intended production. The MOG estimates the likelihood of a given input (e.g. a specific VOT) as the sum of the likelihood that the input was obtained from each of the *K* Gaussians:
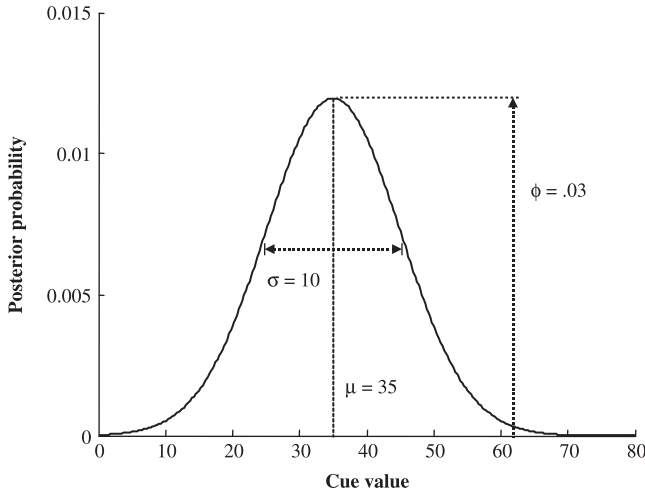
Figure 2 *The parameters of the Gaussian distributions used to represent each speech category. The mean (μ) refers to the location (in cue-space) of the prototype, the SD (σ) refers to the width of the category, and the height (φ) to its frequency (or posterior probability).*

$$M(vot) = \sum_{i=1}^{K} G_i(vot). \tag{2}$$

For example, the probability of obtaining a VOT of 40 ms from a mixture of $K = 2$ Gaussians is the sum of the likelihood that it was generated by either category: the low probability that it arose from the voiced category (e.g. $\mu = 0$ ms of VOT, $\sigma = 10$ ms), plus the higher probability that it arose from a voiceless category (e.g. $\mu = 50$ ms of VOT, $\sigma = 20$ ms). Category membership is simply a matter of determining which Gaussian in the mixture was most likely for a given input.

The MOG approach can fit any continuous cue that forms clusters in the input, and it assumes only that each input category creates a different probability distribution. Speech category learning is simply a matter of estimating the number of categories and their parameters from the input. Multiple contrasts (e.g. voicing and place) can be learned by estimating the parameters of multiple mixtures simultaneously, and Toscano and McMurray (2008) have shown that this framework can be extended to weight multiple cues for the same category. The MOG approach assumes that the psychological representation of such categories is Gaussian. This is not unreasonable, given the prevalence of Gaussian tuning curves in the auditory cortex, and the fact that speech categories exhibit considerable gradiency within categories that takes a more or less Gaussian form (Miller & Volaitis, 1989; McMurray, Tanenhaus & Aslin, 2002).

The MOG approach has been used by De Boer and Kuhl (2003) to demonstrate that a MOG can learn a small number of vowel categories, and that infant-directed speech (Kuhl, Andruski, Chistovich & Chistovich, 1997) can facilitate acquisition. Their model, however, learned with expectation maximization (EM), which employs a complex, multi-stage procedure and estimates parameters after receiving a large *batch* of input. Infants, however, must learn iteratively (i.e. learning occurs after *each* input). In addition, the model was incapable of learning the number of categories needed for a given contrast (the number was specified *a priori*). Because different languages have different numbers of categories, the model must be able to determine this on the basis of the input. Thus, we developed an iterative approach to learning that was simple, plausible, and capable of learning the correct number of categories to fit the native language structure.

The learning algorithm

Our learning algorithm is based on maximum likelihood estimation (MLE) by stochastic gradient descent. MLE is a standard way to estimate the parameters of any function by maximizing the likelihood of the parameters given the data. Gradient descent is a general optimization technique, which virtually all connectionist learning rules (as well as classical learning theory) implement in some way (see Barto, 1995, for an overview). Gradient descent simply adjusts the parameters of the likelihood function (equation 2) along the derivative of the likelihood function (with respect to the parameter). When the derivatives become 0, no further change in the parameters is possible – the function has reached maximum likelihood. This can be a local maxima (for example if the model overgeneralized to a single category), or a global maxima, the best parameter set for the data.

Our model uses gradient descent to adjust μ, σ and ϕ as it encounters individual VOTs (or other cues). This then maximizes the likelihood of the input given the particular parameter set. This allows us to model learning, as it happens, moment by moment (see McMurray, Horst, Toscano & Samuelson, in press, for a theoretical discussion). Thus, our learning rules are as follows:

$$\Delta\phi_i = \eta_\phi \cdot \frac{G_i(vot)}{M(vot)}. \tag{3}$$

After updating ϕ, the vector of all ϕs is normalized to sum to 1 because ϕ represents a prior probability that must be between 0 and 1.

$$\Delta\mu_i = \eta_\mu \cdot \left(\frac{G_i(vot)}{M(vot)} \right)\frac{(vot - \mu_i)}{\sigma_i^2}, \tag{4}$$

$$\Delta\sigma_i = \eta_\sigma \cdot \left(\frac{G_i(vot)}{M(vot)} \right)(\sigma_i^{-3}(vot - \mu_i)^2 - \sigma_i^{-1}). \tag{5}$$

In all three equations, η is a learning-rate parameter, and μ_i, σ_i and ϕ_i represent the parameters of Gaussian i (G_i). As before, $M(vot)$ is the sum of all G_is. These learning rules are applied to each Gaussian after each input. By updating each parameter in small increments (η), the system gradually moves to a set of Gaussians whose parameters are more optimal for the dataset. After sufficient learning, these parameters will typically be truly

Table 1 *Initial parameters for simulations*

Simulation no.	Starting state				Learning rates			Input language		
	K	σ (ms of VOT)	μ (ms of VOT)	ϕ	η_μ	η_σ	η_ϕ	μ	σ	ϕ
1a: Learning	25	2	Random	$1/K$	1	1	.0001	[0 50]	[5 15]	[.5 .5]
1b: Learning + competition	25	2	Random	$1/K$	1	1	.0001	[0 50]	[5 15]	[.5 .5]
2: Pruning	10	5	Random	$1/K$	.2	.2	.0005	[0 50]	[5 10]	[.5 .5]
3: Enhancement	10	20	Random	$1/K$	.2	.2	.0005	[0 50]	[20 20]	[.5 .5]
4: Sparseness	4–50	1–60	Random	$1/K$	.2	.2	.004	[0 50]	[12 12]	[.5 .5]

optimal (a global minimum). However, occasionally the model will converge on a local minimum in which the parameters are not optimal, but cannot be further improved. In all of the simulations that were run, a local minimum was seen only when the model incorrectly estimated the number of Gaussians with non-zero ϕs: either overgeneralizing (too few Gaussians) or undergeneralizing (too many).

The model was trained as follows. First, an array of K Gaussians is randomly generated to serve as the initial state. K is relatively high (e.g. 10–20) as the model does not know how many categories (Gaussians) it will need. The μs of these Gaussians are randomly selected, σs are set to a small constant value,[1] and ϕs to $1/K$. After initializing the model, it is exposed to a set of inputs. On each generation, a single value of a speech cue is selected. For these simulations, VOT was generally used, but this is arbitrary – a MOG model can be applied to any continuous cue. These cue-values are randomly generated from a bimodal (two-category) Gaussian distribution (whose means and standard deviations are based on the means and standard deviations reported by speech production measurements such as Lisker & Abramson, 1964). After this cue-value is selected, the three parameters (ϕ, μ and σ) of each of the K Gaussians are adjusted according to the learning rules in Equations 3–5. This is repeated until the model converges on a solution (the parameters reach asymptote), usually after several thousand iterations.[2]

Tests of the model

Simulation 1: Statistical learning requires competition

The first simulations determined if these learning rules were sufficient for speech categories to be learnt. One hundred models were initialized (parameters in Table 1, Simulation 1) and trained for 100,000 generations on a random sampling from a dataset based on Lisker and Abramson's (1964) estimates of English VOTs. We then determined (1) if the model converged on the correct number of categories (i.e. if the frequency parameter,

ϕ, was much greater than .01 for the correct number of Gaussians), and (2) if those categories approximated the training distribution (had correct values for μ and σ).

None of the models converged on a reasonable solution, averaging 11.9 ($SD = 1.8$) active Gaussians, and no model reached the true two-category solution. Thus, although the model suppressed some Gaussians (K was 25), it never arrived at the correct number. Nonetheless, the model could have approximated the training distribution across a *set* of Gaussians (a distributed representation). To test this, each active Gaussian was categorized as belonging to either the voiced or the voiceless category (whichever it was closest to), and the parameters of the corresponding sets of Gaussians were analysed to determine if they collectively approximated the input. First, we considered whether ϕ was estimated correctly. Because multiple Gaussians were above-threshold for each category, the sum of their ϕs should be .5 within a single category. However, in only 57 of the 100 models was the sum of the ϕs in a single category between .45 and .55. μ and σ were equally inaccurate. The average RMS difference of each Gaussian's μ from the closest input category was 6.14 ms of VOT ($SD = .9$), and for σ this value was 3.68 ms ($SD = .9$). Thus, even a representation distributed across multiple Gaussians did not match the input distribution.

To improve the model, a simple change was introduced: winner-take-all competition. The model used the same learning rules, with the exception that ϕ was changed only for the single Gaussian that had the highest likelihood for the current input. This is psychologically plausible, as only the frequency of one category should be adjusted for a given input. Computationally, it approximates Rumelhart and Zipser's (1986) competitive learning, which McMurray and Spivey (1999) applied to bimodal Gaussian distributions.

We implemented 100 new models and found that 97 arrived at the correct two-category solution. For these 97 models, ϕ was always between .45 and .55 for the two categories, μ averaged .52 ms of VOT ($SD = .28$) from that of the closest input category, and σ averaged .69 ms from that of the nearest input category ($SD = .63$). Thus, competition allowed the MOG to learn the correct number of categories and to align them nearly perfectly with the input.
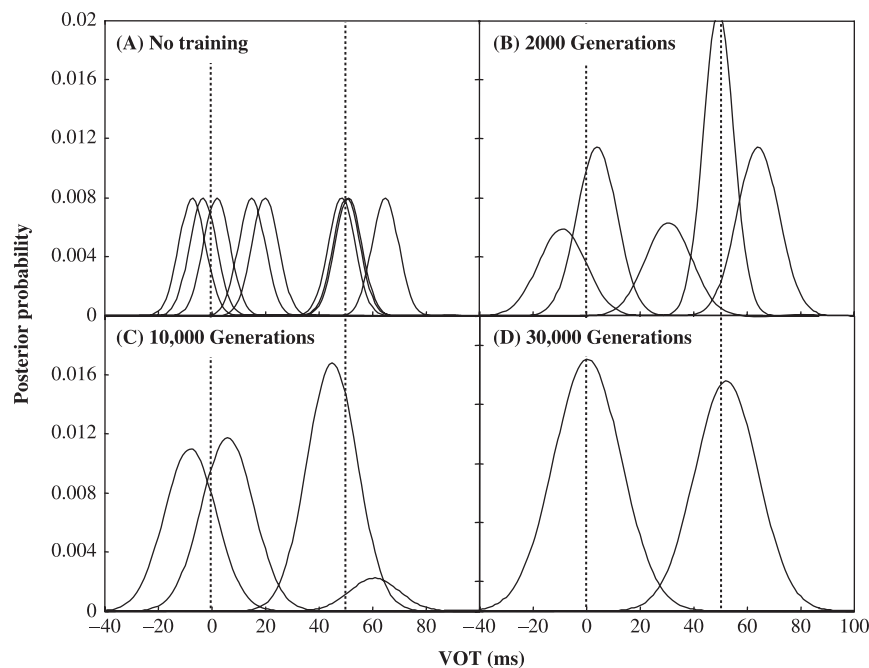
The discreteness of winner-take-all competition, however, may oversimplify what is probably a fundamentally continuous process. Thus, we attempted three additional competition schemes: simple linear normalization, quadratic normalization, and the softmax function (with and without

---

[1] Although $\sigma_{initial}$ is arbitrarily set by the experimenter, Simulation 4 tested success on $\sigma_{initial}$ ranging from 1 ms of VOT to 50, and found that σs between 3 and 25 were largely successful (Figure 5A).

[2] The MOG described here was implemented in MATLAB (code is available from the first author).

Table 2 *Average number of categories after 100,000 generations of learning as a function of competition rules. Numbers in parentheses are standard deviations. Numbers on the second line refer to the proportion (of 100 simulations) that correctly extracted two categories*

	Implementation of ϕ (frequency)			
	ϕ (original)	$(\phi_i)/\left(\sum_{j\in k}\phi_j\right)$ (normalized)	$(\phi_i^2)/\left(\sum_{j\in k}\phi_j^2\right)$ (quadratic normalized)	$(\exp(T\phi_i))/\left(\sum_{j\in k}\exp(T\phi_j)\right)$ (softmax)
No competition	11.9 (1.8)	10.6 (2.3)	3.8 (1.3)	3.5 (1.5)
	0%	0%	16%	33%
Winner take all	1.97 (.17)	1.99 (.1)	1.93 (.26)	1.89 (.31)
	97%	99%	93%	89%

Figure 3 *A single mixture of Gaussians (MOG) model over the course of training on a distribution with means of 0 and 50 and equal standard deviations of 15. Dashed vertical lines represent the means of the two training categories. (A) With no input, all K (8) Gaussians are equally likely. (B and C) After a few thousand exposures, the model suppresses some of the unnecessary Gaussians, until (D) by the end of training, only the two correct Gaussians remain.*

winner-take-all). As Table 2 shows, without competition, quadratic normalization and softmax outperformed the original implementation. However, when competition was included, these two schemes offered no additional benefit over the original model. Thus, this form of statistical learning is only successful with competition, and, of the competition methods examined, winner-take-all seems to yield the best performance.

A number of unsupervised connectionist architectures show the same property. Competitive Hebbian learning (e.g. Rumelhart & Zipser, 1986) uses winner-take-all competition; Hebbian normalized recurrence uses quadratic normalization and cannot learn speech categories without it (McMurray *et al.*, in press); and self-organizing feature maps (Guenther & Gjaja, 1992) employ a topographic excitation/inhibition rule. These architectures buttress the current point: competition is required for distributional category learning.

Modelling the developmental timecourse

Using the hybrid model (learning + competition), we now ask whether it accounts for the developmental timecourse of phonetic category formation. Figure 3 shows a characteristic run of the model. Over the course of development, unnecessary Gaussians are eliminated and the remaining ones adjust to fit the input from the training distribution. Thus, at 2,000 training generations (panel B), the model has a large number of categories that are not aligned to the training data. At this point, any two VOTs are likely to fall under different categories and be easily discriminable. However, after 30,000 generations (Figure 3D), the model successfully represents the input, and many (within-category) contrasts will fall under the same Gaussian and be indiscriminable.

This simplistic analysis assumes that infants only discriminate tokens that fall completely into different categories. Early in development, however, inputs may
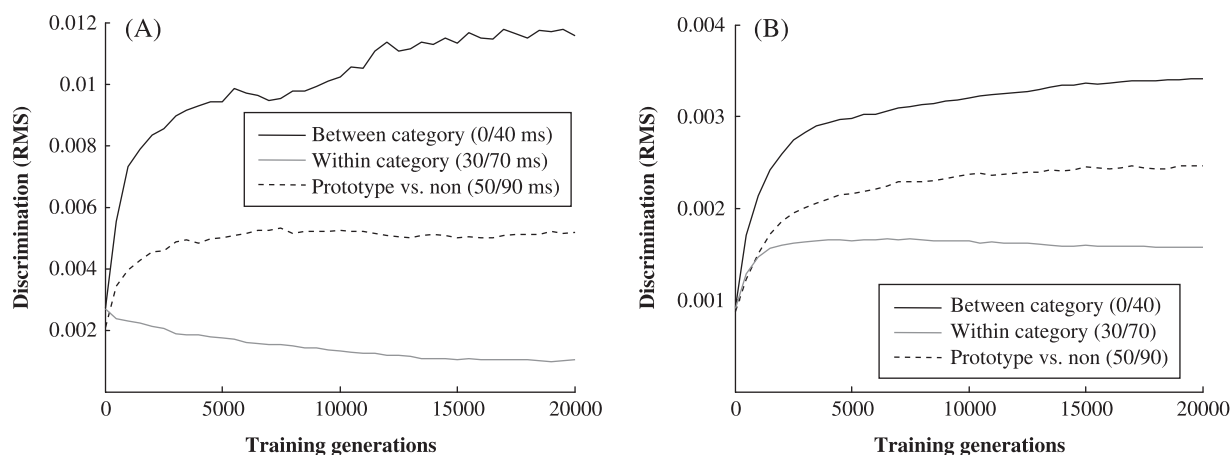
Figure 4 *Changes in the RMS discrimination metric over the course of training. Models were tested every 500 generations on three comparisons: discrimination between 0- and 40-ms tokens that crossed the category boundaries (black lines); discrimination between 30- and 70-ms tokens in the same category (grey lines); and discrimination between a prototypical (50 ms) and a non-prototypical (90 ms) exemplar (dashed lines). (A) The average of 30 models that started with small σs and were given as input the distribution of voice onset time (VOT) in English. (B) The average of 30 models that started with large σs and were given as input the highly overlapping distribution of exemplars that simulate English fricatives.*

fall under a number of overlapping categories. Discrimination could occur for differences between any of these Gaussians. We developed a discrimination measure to account for this. Each of the two VOTs that were to be compared was converted to a *K*-length vector of the probabilities of each Gaussian (category). The RMS distance of these vectors was used to compare the two VOTs in 'category-space'. A small RMS arises if baseline and comparison stimuli are represented with a similar set of categories. On the other hand, a large RMS indicates largely different sets of categories.

Simulation 2: Pruning

Many phonetic dimensions exhibit an overgeneration/pruning pattern over the course of development: infants are initially sensitive to a wide range of phonetic contrasts, and lose sensitivity to contrasts that are not used (e.g. Werker & Tees, 1984). To model this, we exposed 30 models to the Lisker and Abramson (1964) distribution of English VOTs for 30,000 epochs using the parameters described in Table 1 (Simulation 2). They were tested every 500 epochs on three VOT contrasts: 0 vs. 40 ms (tokens from opposite categories), 30 vs. 70 ms (a within-category difference that should be lost), and 50 vs. 90 ms (still within-category, but the difference is between a prototypical and non-prototypical token; for example, Miller & Eimas, 1996).

Figure 4A shows the discrimination performance of the model (RMS) over the course of training. Initially, the model is equally good at all three contrasts. Over the course of training, between-category discriminability increases, whereas within-category discrimination is lost. Discriminability between prototype and non-prototype distinctions also increases (as the model extracts the

structure of the category), but never approaches between-category discriminability. Thus, the model starts with some ability to discriminate all three contrasts and loses the ones it does not need.

Simulation 3: Enhancement

A small number of speech contrasts (e.g. s/z and f/θ, Eilers & Minifie, 1975; Eilers, Wilson & Moore, 1977) show developmental enhancement: infants initially lack the ability to discriminate a meaningful speech contrast but develop it later.

To simplify this problem to a single dimension, we assumed that the underlying cue for fricative discrimination was sensitivity to the spectral mean of the frication noise. Because these spectral-mean detectors have Gaussian tuning curves, and many frequencies are present at once for a fricative, the starting categories ($\sigma_{initial}$) and statistical distributions are quite broad. Thirty additional simulations were run using a hypothetical training distribution based on these estimated spectral means. These simulations were identical to the prior ones for VOT except that the model's σs ($\sigma_{initial}$) started out broad ($\sigma = 20$), and the distributions of the input were highly overlapping (Table 1, Simulation 3).

Again, the model started with relatively equal abilities to discriminate the three contrasts (Figure 4B). As in the previous simulation, the between-category contrast was enhanced over the course of learning, as was the contrast between the prototypical and non-prototypical exemplars. In addition, the within-category contrast was *enhanced* slightly. Although this seems counterintuitive, if we assume that within-category contrasts are difficult to discriminate in adulthood, this upward trend would imply that everything is indiscriminable early.

Simulation 4: Sparseness

The MOG model can account for the developmental trends in speech categorization along several different acoustic/phonetic dimensions. However, it also provides novel insights about development. One non-obvious implication of this model is that infants do not learn category *boundaries*; rather, they learn the distribution of exemplars that define a category. Because each category is defined independently of any others, they are not required to completely encompass the phonetic space, and there may be regions of the phonetic space that are not mapped to any category (a gap).

McMurray and Aslin (2005) provided evidence that is suggestive of such a sparse representation. After being exposed to a series of words with syllable-initial VOTs near 3–4 ms, infants discriminated them from tokens with VOTs near 12 ms (and these medial tokens were not discriminated from 40 ms). This could be explained by a category boundary between 3–4 and 12 ms, although there is no evidence for such a boundary in any language or age group. Alternatively, there may be no category at all in the middle region of the VOT dimension. To test this conjecture, a series of simulations evaluated the amount of input space that was uncategorized over the course of learning. Because this sparseness value is likely to be related to the number of categories available to fill this space (K), and to the width of the initial categories ($\sigma_{initial}$), a range of 21 σs (from 1 to 60) and 13 Ks (from 4 to 50) were selected. Fourteen models were trained with each combination of these parameters (Table 1, Simulation 4).

As in the prior simulations, the number of categories learned (i.e. $\phi > .01$ after training) was used to measure success. Both K and $\sigma_{initial}$ were related to success. Although the model was successful over a large range of σs, it tended to fail when $\sigma_{initial}$ exceeded 25 ms of VOT (Figure 5A), half the distance between the category means (50 ms). In a sense, it was easier for the model to work from small to large categories than to divide initially large categories. Large Ks could mitigate this effect, but not eliminate it: even with $K = 50$, no model was able to learn when $\sigma_{initial}$ was greater than 40.

To estimate the amount of input space that the model left uncategorized between 0 and 50 ms (the two proto-type VOTs), we computed a sparseness coefficient (SC). At each VOT, the posterior probabilities of each of the K Gaussians were computed. If any of these was higher than 10% of the maximum posterior,[3] that point was said to have been categorized; otherwise, it was uncategorized. The SC was the percentage of these 51 VOTs (between 0 and 50 ms) that were left uncategorized.

Figure 5B displays the SC as a function of training epochs and starting σ. Not surprisingly, models starting with large σs were not sparse – these wide categories encompassed most of the cue-space. Also not surprisingly,

[3] 25% was also tried as a criterion with similar results.

small initial σs (1–4) yielded early sparseness that gradually decreased. Interestingly, however, medium σs (5–20) showed an initial lack of sparseness, followed by a rapid increase between 250 and 1000 generations, and finally a decrease to complete representations. Most of the models with large σs (71%) failed at learning the input (compared with 98% and 76% for medium and small σs, respectively), implying that sparseness arises naturally for most of the successful starting states.

The same pattern was seen with respect to K (Figure 5C). Ks greater than 20 had quite complete representations initially, whereas medium and small Ks started out sparser. However, all Ks showed an increase (and decrease) in sparseness between 250 and 3500 generations. Moreover, the largest increases occurred for the largest values of K. As before, this increase in sparseness appeared to be related to success: large Ks were associated with the greatest success rate (83%) compared with medium (79%) and small (69%) Ks. Thus, again, optimal starting parameters led (developmentally) to sparse representations of the input, even where the optimal K led to an initially complete representation.

A hierarchical logistic regression was used to determine if sparse representations led to successful learning. Success was regressed against K, $\sigma_{initial}$ and SC (between 250 and 3000 generations). In the first step of the regression, significant main effects of both $\sigma_{initial}$ ($B = -.105$, Wald(1) = 815.2, $p < .001$) and K ($B = .04$, Wald(1) = 91.2, $p < .001$) were found. Here, $\sigma_{initial}$ was inversely correlated with success (lower σs led to greater success), whereas K was positively correlated. In the second step, the $K \times \sigma$ interaction was significant ($B = -.002$, Wald(1) = 48.4, $p < .001$): higher Ks allowed the model to overcome larger σs. In the third step, SC was added and was highly significant, over and above the other two factors ($\chi^2_{change}(1) = 67.4$, $p < .0001$; $B = 14.2$, Wald(1) = 43.5, $p < .0001$). Thus, although K and σ influenced sparseness, SC had a positive effect on success beyond that of K and σ (and their interaction). Models that arrived at sparse representations were more likely to succeed after further input than those that did not.

Conclusions

The MOG approach to the statistical learning of speech categories highlights a number of important points. First, statistical learning is insufficient to accomplish the task: competition of some kind is required. Competition is a property of many models in many domains, including unsupervised connectionist architectures and models of adult word recognition (McClelland & Elman, 1986; Luce & Pisoni, 1998). Moreover, once competition is incorporated into the model, it accounts for both developmental trajectories observed empirically: overgeneration/pruning and enhancement. Here, the specific trajectory does not arise from differences in developmental mechanisms, but rather from differences
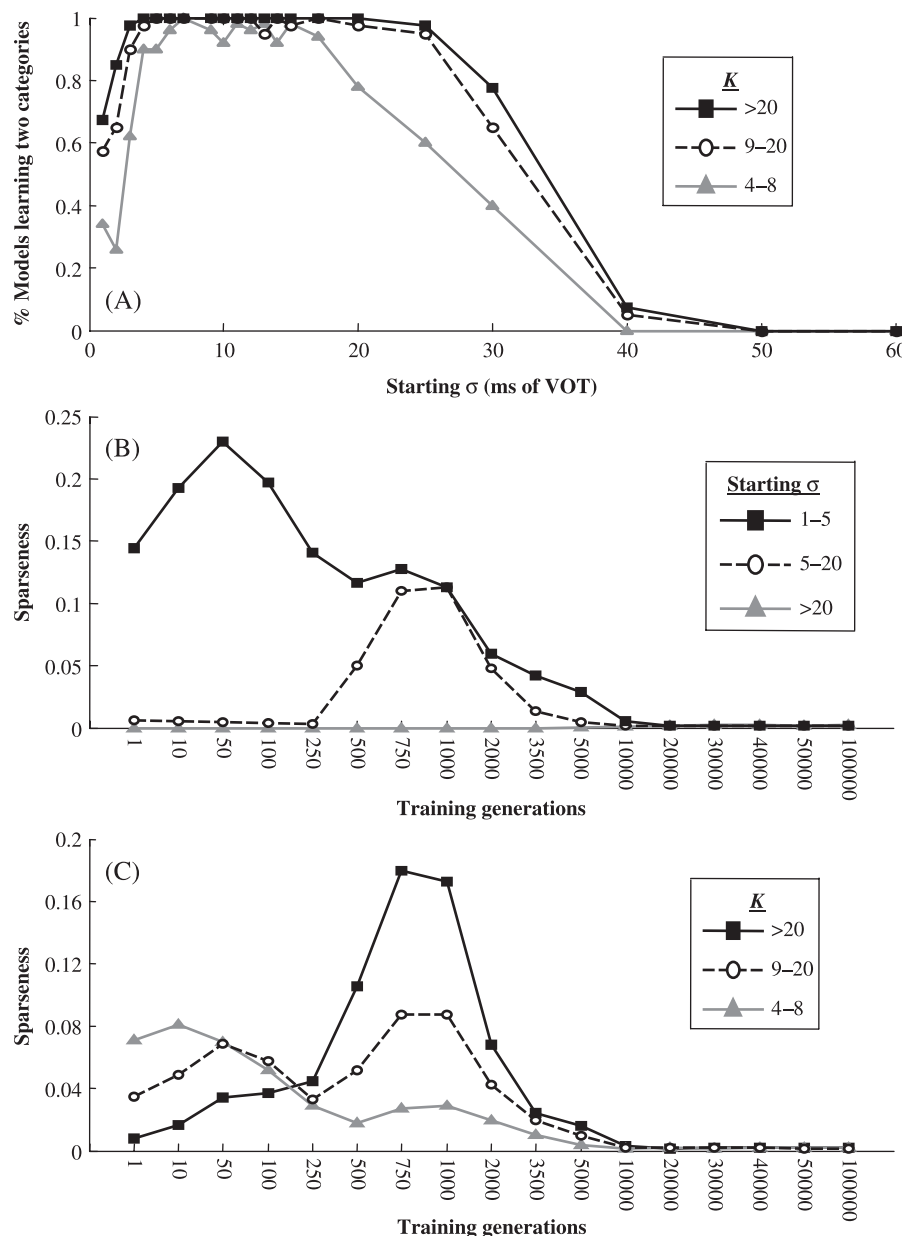
Figure 5 *Results of Simulation number 4: Sparseness. (A) The effect of starting σ and K on the probability of success (learning the correct two-category solution). (B) Sparseness coefficient, SC (the proportion of the input space not mapped onto a category) over the course of training for small, medium and large initial σs. (C) SC over the course of training for small, medium and large values of K.*

in how the cues are perceived and in their statistical distributions.

An alternative to the competition/distributional learning account presented here might make sole use of the counts of each VOT. Such an alternative would simply track the frequency of individual VOTs, perhaps recording these counts by warping perceptual space. However, for this set of frequency statistics to create a set of categories, a decision criterion must be employed. This or any other decision process would invariably involve competition. As a result, we agree with Remez (2005) – simple counts of token frequency may not be sufficient for category learning. However, across a range of architectures (the MOG as well

as the connectionist architectures discussed), competition can transform these counts into useful categories.

Second, the MOG model implies that infants are not learning phonological distinctions (e.g. voicing), but rather that the process of category acquisition is one in which isolated regions of these dimensions are gradually grouped together. Categories are independent of one another and do not need to completely encompass a given dimension (at least early in learning). Our simulations demonstrate that even models that do not start out sparse go through a sparse stage and that sparseness is correlated with later success. By not categorizing certain regions of the input (typically the more ambiguous regions), the

model is, in a sense, waiting for more data before committing to a mutually exclusive category structure.[4]

Finally, the implications of sparseness suggest a different understanding of classic data concerning the seemingly counterintuitive ability of young infants to discriminate non-native phonetic contrasts. Colloquially, this ability is often described as infants 'having' non-native categories. However, the MOG model suggests that discrimination could also occur when one input is categorized and one falls into a sparse region of the space (no category).

This computational work provides further evidence for the plausibility of unsupervised learning of speech categories via a statistical learning mechanism. Our implementation suggests that statistical learning alone is not sufficient for robust learning. However, when combined with another core mechanism (competition), the MOG yields not only successful data-driven learning that approximates the developmental timecourse, but also novel insights about the sparse nature of early speech categories.

Acknowledgements

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.

Barto, A.G. (1995). Learning as hill-climbing in weight space. In M. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 531–533). Cambridge, MA: MIT Press.

Bertoncini, J., Bigeljac–Babic, R., Blumstein, S.E., & Mehler, J. (1987). Discrimination in neonates of very short CV's. *Journal of the Acoustical Society of America*, **82** (1), 31–37.

Best, C.T., McRoberts, G.W., & Sithole, N.M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, **14** (3), 345–360.

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

Eilers, R., & Minifie, F. (1975). Fricative discrimination in early infancy. *Journal of Speech and Hearing Research*, **18** (1), 158–167.

Eilers, R., Wilson, W., & Moore, J. (1977). Developmental changes in speech discrimination in infants. *Perception & Psychophysics*, **16**, 513–521.

Eimas, P.D., Siqueland, E.R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, **171**, 303–306.

Elman, J., & Zipser, D. (1988). Learning the hidden structure of speech. *Journal of the Acoustical Society of America*, **83** (4), 1615–1626.

Espy-Wilson, C.Y. (1992). Acoustic measures for linguistic features distinguishing the semi-vowels /w j r l/ in American English. *Journal of the Acoustical Society of America*, **92** (1), 736–757.

Guenther, F., & Gjaja, M. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America*, **100**, 1111–1112.

Hillenbrand, J.M., Getty, L., Clark, M.J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, **97** (5), 3099–3111.

Jusczyk, P. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.

Kuhl, P.K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, **5**, 831–843.

Kuhl, P.K., Andruski, J.E., Chistovich, I., & Chistovich, L. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, **277** (5326), 684–686.

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

McCandliss, B.D., Fiez, J.A., Protopapas, A., Conway, M., & McClelland, J.L. (2002). Success and failure in teaching the [r]–[l] contrast to Japanese adults: tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective, & Behavioral Neuroscience*, **2**, 89–108.

McClelland, J., & Elman, J. (1986). The TRACE model of speech perception. *Cognitive Psychology*, **18** (1), 1–86.

McMurray, B., & Aslin, R.N. (2005). Infants are sensitive to within-category variation in speech perception. *Cognition*, **95** (2), B15–B26.

McMurray, B., & Spivey, M.J. (1999). The categorical perception of consonants: the interaction of learning and processing. *Proceedings of the Chicago Linguistics Society*, **35**, 205–219.

McMurray, B., Tanenhaus, M.K., & Aslin, R.N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, **86** (2), B33–B42.

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

---

[4] In the extreme, the sparseness approach could be interpreted as a sort of null category encompassing any uncategorized region along the cue-dimension. Under this view it is possible that infants would treat sparse regions of VOT *between* the two voicing categories as members of the same category as a sparse region *outside* the categories (e.g. a very long VOT). However, it is accepted that phonetic discrimination in adults is a function of both continuous stimulus differences and discrete category differences (Pisoni & Tash, 1974). Thus, it is likely that infant phonetic discrimination can take advantage of both continuous differences, the emerging phonetic categories (or null categories), and, as we discussed, the marginal activations of neighbouring categories. This would result in infants discriminating tokens from two sparse regions (that differed physically).

Maye, J., & Gerken, L. (2000). Learning phonemes without minimal pairs. *The Proceedings of the Boston University Conference on Language Development*, **24**, 522–533.

Maye, J., Werker, J.F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, **82**, 101–111.

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

Miller, J.L. (1997). Internal structure of phonetic categories. *Language and Cognitive Processes*, **12**, 865–869.

Miller, J.L., & Eimas, P.D. (1996). Internal structure of voicing categories in early infancy. *Perception & Psychophysics*, **58** (8), 1157–1167.

Miller, J.L., & Volaitis, L.E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics*, **46** (6), 505–512.

Murphy, G. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Nakisa, R., & Plunkett, K. (1998). Evolution of a rapidly learned representation for speech. *Language and Cognitive Processes*, **13** (2&3), 105–127.

Peterson, G.E., & Barney, H. (1951). Control methods used in the study of vowels. *Journal of the Acoustical Society of America*, **24** (2), 175–184.

Remez, R. (2005). Perceptual organization of speech. In D.B. Pisoni & R.E. Remez (Eds.), *The handbook of speech perception* (pp. 28–50). Oxford: Blackwell Publishing.

Rumelhart, D.E., & Zipser, D. (1986). Feature discovery by competitive learning. In D.E. Rumelhart & J. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 151–193). Cambridge, MA: MIT Press.

Titterington, D.M., Smith, A.F., & Makov, U.E. (1985). *Statistical analysis of finite mixture distributions*. New York: Wiley.

Toscano, J., & McMurray, B. (2008). Using the distributional statistics of speech sounds for weighting and integrating acoustic cues. In B.C. Love, K. McRae, & V.M. Sloutsky (Eds.), *Proceedings of the Cognitive Science Society* (pp. 433–439). Austin, TX: Cognitive Science Society.

Werker, J.F., & Curtin, S. (2005). PRIMIR: a developmental framework of infant speech processing. *Language Learning and Development*, **1** (2), 197–234.

Werker, J.F., & Lalonde, C.F. (1988). Cross-language speech perception: initial capabilities and developmental change. *Developmental Psychology*, **24** (5), 672–683.

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