Modeling the categorical perception of speech sounds: A step toward biological plausibility

NELLI H. SALMINEN, HANNU TIITINEN, AND PATRICK J. C. MAY
Helsinki University of Technology, Helsinki, Finland

Our native language has a lifelong effect on how we perceive speech sounds. Behaviorally, this is manifested as categorical perception, but the neural mechanisms underlying this phenomenon are still unknown. Here, we constructed a computational model of categorical perception, following principles consistent with infant speech learning. A self-organizing network was exposed to a statistical distribution of speech input presented as neural activity patterns of the auditory periphery, resembling the way sound arrives to the human brain. In the resulting neural map, categorical perception emerges from most single neurons of the model being maximally activated by prototypical speech sounds, while the largest variability in activity is produced at category boundaries. Consequently, regions in the vicinity of prototypes become perceptually compressed, and regions at category boundaries become expanded. Thus, the present study offers a unifying framework for explaining the neural basis of the warping of perceptual space associated with categorical perception.

The speech categories of a native language have a lasting effect on the organization of the auditory system in humans (Kuhl, 2000). This is evident even in the perception of isolated vowels and consonants, which, in themselves, carry no linguistic meaning. Acoustically equal distances between speech sounds are perceived as unequal, and this warping of the perceptual space is related to the native language of the subject. This has been described as either expansions at the boundaries of phonetic categories (categorical perception; Liberman, Safford Harris, Hoffman, & Griffith, 1957) or compressions within category limits (the perceptual magnet effect; Kuhl, 1991; Samuel, 1982). In the case of categorical perception, two speech sounds can be told apart with greater ease when the sounds belong to different phonetic categories than when they belong to the same category (Eimas, 1963; Liberman et al., 1957; Macmillan, Goldberg, & Braida, 1988; Pisoni, 1973; Repp, 1984; Wood, 1976). The compression around the prototypical speech sounds, in turn, has been explained in terms of perceptual magnets (Aaltonen, Eerola, Hellström, Uusipaikka, & Lang, 1997; Frieda, Walley, Flege, & Sloane, 1999; Iverson & Kuhl, 1995, 2000; Kuhl, 1991). The prototypes of the phonemes of the subject's native language distort the perceptual space around them by pulling the percepts of nearby sounds toward themselves. Both of these theoretical frameworks, however, leave open the question of what the neuronal and computational mechanisms underlying warped perception could be.

Attempts have been made to explain warped perception in terms of self-organizing neural maps (Bauer, Der, & Herrman, 1996; Guenther & Gjaja, 1996). These employ

the biologically plausible Hebbian learning rule and require no explicit training signals on the category membership of the stimuli (Sejnowski, 1999; Yuste & Sur, 1999). Warping has been suggested to arise from the distribution of neuronal resources according to two alternatives, both of which can emerge from self-organization. The first alternative, proposed by Guenther and Gjaja, is that more neurons are preferentially activated by the prototypes of the speech categories than by other speech sounds, including those at the category boundaries. This neuronal overrepresentation was offered as an explanation for perceptual warping. The percepts of the model are calculated in terms of maximal tuning and population vectors. Each neuron is thought to "vote" for the perception of its preferred speech sound, and in this process, the majority preferring the prototypical instances dominate the result whenever the speech sound is close to a prototype. Due to this dominance, a speech sound in the proximity of the prototype is perceived to be more similar to the prototype than would be expected on the basis of the physical distance between the two. However, in this framework, the relation between the distribution of neuronal resources and the level of perceptual performance appears to be problematic: The category boundaries are more sparsely represented than the prototypes, and, as was pointed out by Guenther, Husain, Cohen, and Shinn-Cunningham (1999) and Guenther, Nieto-Castanon, Ghosh, and Tourville (2004), it seems counterintuitive that behavioral discrimination would improve as the amount of neuronal resources encoding the stimuli decreases.

The second, alternative explanation relates the improved discriminability at the category boundaries to an

increased number of neurons encoding these boundaries (Bauer et al., 1996; Guenther et al., 1999). Bauer et al. formulated a self-organizing learning rule through which such a representation could be achieved. In this model, the category boundaries are represented densely with a large population of model neurons, and better perceptual abilities are assumed to arise from this overrepresentation (Bauer et al., 1996). The prototypical instances, in turn, are represented with a small number of neurons, and correspondingly, discriminability is poor for these stimuli. In this work, the better discriminability was assumed to arise from neuronal overrepresentation, but no analyses were presented to verify this. Such a relation between neuronal overrepresentation and increased discriminability has, however, received support from experimental work. Hemodynamic recordings show that when subjects undergo pitch discrimination training, an increase in discriminability is accompanied by a higher level of activity for the stimuli at the training-induced category boundary (Guenther et al., 1999; Guenther et al., 2004). However, from the perspective of speech perception, this type of tuning seems problematic. It would mean that the neurons in the brain of a developing infant would need to become selective to seldom-heard speech sounds, instead of frequently occurring ones. Thus, the human brain would be specialized in representing untypical sounds of the environment, instead of typical ones.

Computational models of categorical perception have so far focused on the processing of abstract and simplified input patterns that bear little resemblance to speech sounds (Anderson, Silverstein, Ritz, & Jones, 1977; Bauer et al., 1996; Goldstone, Steyvers, & Larimer, 1996; Harnad, Hanson, & Lubin, 1991, 1995; Kruschke, 1992). Even in studies specifically modeling speech processing, the input has been highly abstracted: Typically, the readily extracted values of a physical sound feature, such as the formant frequencies, have been used (Guenther & Gjaja, 1996; Vallabha & McClelland, 2007). This, however, is not how the human auditory nervous system receives and represents speech. Instead, the input arriving through the auditory nerve is a tonotopically organized pattern of neural activity. The possible downside of the use of abstracted speech input in modeling was demonstrated by Damper and Harnad (2000), who trained neural networks to categorize voice-onset-time (VOT) consonant continuums. Rather than constructing simplified input on the basis of physical parameters of the input (e.g., the VOT in milliseconds), they utilized the original speech sound waveforms preprocessed with a model of the periphery of the auditory nervous system. As a result, the networks placed the category limits on the same values of VOT as do human and animal subjects, even though the networks were not presented with any information on the location of the boundary during training. A model utilizing the more abstract input coding would not have been able to reveal this natural boundary in the input continuum. Thus, it seems that even though the speech inputs are at equal distances according to some physical measure, such as VOT or formant frequency, these distances are no longer equal in the neuronal representation utilized by the auditory system.

Further simplifications have been made in previous computational work on categorical perception by providing models with explicit input on category membership (Goldstone et al., 1996; Harnad et al., 1991, 1995; Kruschke, 1992) or by utilizing only the prototypical instances of the categories (Anderson et al., 1977; Damper & Harnad, 2000). In contrast, humans do not learn speech categories on the basis of explicit category labeling or just by listening to prototypical instances. Instead, they learn from sounds containing considerable variability, originating from different speakers and different contexts (Kuhl et al., 1997; Peterson & Barney, 1952). In the case of vowels, a wide range of combinations of formant frequencies occur in the input, and the instances of different categories do not fall within some strict limits determined by their acoustical properties. Despite this large variability, infants are capable of extracting information on the speech categories of their native language from the ambient language input within the first 6 months of life (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992), and this learning seems to be based on the distributional properties of the speech input (Kuhl et al., 1997; Maye, Werker, & Gerken, 2002).

In short, previous modeling studies have offered conflicting suggestions as to the neuronal origins of warped perception. On the one hand, the neuronal resources being dedicated mainly to the prototypical speech sounds could explain the compressed perception around category prototypes, but this leaves open the question of how increased perceptual abilities could arise from a smaller neuronal population encoding the category boundaries. On the other hand, a larger population being dedicated to category boundaries seems counterintuitive: It would mean that neuronal resources are dedicated to untypical speech sounds that poorly represent the speech categories of the native language. The aim of this work was to resolve these inconsistencies with a computational model that follows biologically plausible principles both in the speech input and in the model dynamics. A set of speech sounds was preprocessed with a model of the peripheral auditory nervous system. This input, following a statistical distribution describing the occurrence of sounds in ambient speech input, was presented to a self-organizing neuronal network. The resulting neuronal representation of speech sounds was then studied to explore the neuronal mechanisms underlying perceptual warping. As an extension to previous work (Guenther & Gjaja, 1996), we performed analyses on the tuning of single model neurons. This allowed us to identify the stimulus maximally exciting each neuron, as well as the response profile of each neuron across the whole stimulus space.

METHOD

Speech Stimuli

A set of vowel sounds was used as input material. Vowels represent a case for which perceptual warping has been found both at the category boundary (Pisoni, 1973) and at the category center (Aaltonen et al., 1997; Frieda et al., 1999; Iverson & Kuhl, 1995, 2000; Kuhl, 1991). The iden-

tity of vowels is determined by their formant frequency values (Fant, 1970). Thus, by systematically varying these frequencies, one can generate continuums of vowel sounds where the stimuli are evenly spaced according to their physical features (e.g., Iverson & Kuhl, 1995, 2000; Kuhl, 1991). These stimuli can then be used in probing the warping of the perceptual space by studying how the perceptual distances vary across the stimulus space and how this variation relates to the native speech categories.

Vowels were produced with a MATLAB toolbox (cobweb.ecn.purdue.edu/~malcolm/interval/1998-010) that first synthesizes the vocal fold pulses with a specific fundamental frequency and then filters this signal to amplify the desired formant frequencies (Figure 1A). A 2-D input space was defined in terms of the two lowest formant frequencies, and vowel sounds were synthesized at equal steps in the mel scale. This is a psychophysical scale of sound frequency that is used in the study of perceptual distances between speech stimuli (Iverson & Kuhl, 1995, 2000; Kuhl, 1991), and its relationship with the hertz scale is given by

$$f_{\text{mel}} = 1127.01048 * \log(1 + f_{\text{Hz}}/700).$$

The vowel space here spanned 350–800 mel in the first formant ($F1$) and 680–1,700 mel in the second formant ($F2$) frequency in steps of 5 mel. $F1$ – $F2$ pairs in which $F1 > F2 - 200$ mel were excluded in order to avoid pairs in which the formant frequencies were unrealistically close to each other or where $F1$ was higher than $F2$. Further variation to the input material was introduced by adjusting the fundamental frequency of the vowels from 130 to 170 mel in 5-mel steps. Thus, the input material consisted of 137,000 vowel sounds altogether.

Preprocessing With a Model of the Auditory Periphery

The synthesized vowel sounds were preprocessed with a model of the auditory periphery. This first simulates the vibrations of the basilar membrane and then translates these into neural activity resembling that found in the auditory nerve and in the cochlear nucleus (Figures 1B and 1C). Importantly, this activity is expressed as activation patterns in tonotopically organized frequency channels, which is the main principle for representing sound in all parts of the auditory system, including the auditory cortex.

In producing peripheral activity patterns, the first stages of the auditory image model of Patterson, Allerhand, and Giguère (1995; www.pdn.cam.ac.uk/groups/cnbh/aimmanual) were utilized, with a model of 64 frequency channels spanning the 320- to 1,730-mel region. The time-averaged output obtained from the periphery model (Figure 1D) was scaled, while preserving the original structure of the activity patterns. For each frequency channel, the mean activity was set to zero by subtracting the average across all vowels (Figure 1E). This was necessary because, if the mean deviated from zero, network neurons, when allowed to self-organize, would cluster around the average input pattern. Thereafter, the activities were scaled for reporting purposes, with the maximum absolute value of activity calculated across all vowels and frequency channels.

Statistical Distribution of Input Presentation

The speech categories were implicitly modeled by presenting prototypical sounds more often than the others during neural map organization. For this purpose, we used

Figure 1. Construction of the speech input as tonotopically organized neural activity patterns. (A) First, a vowel sound was synthesized. (B) This waveform was then fed to a model of the auditory periphery, which produced a tonotopically organized neural activity pattern. (C) This representation consisted of the activity of 64 frequency channels and was time-averaged. (D and E) Examples of the inputs are presented corresponding to four prototypical speech sounds before and after scaling. The formant frequencies are visible as peaks in activity.

2-D Gaussian probability distributions in the $F1$ – $F2$ space centered at the prototypes of the speech categories of the modeled language environment (see Figure 3A). The Gaussian distributions were chosen on the basis of previous research demonstrating their usefulness in modeling naturally occurring distributions of vowel sounds (Valabha, McClelland, Pons, Werker, & Amano, 2007). Four speech categories, corresponding roughly to the Finnish vowels /o/, /e/, /u/, and /i/, were used. Accordingly, the distributions were centered at $F1$ – $F2$ frequencies of 720–950, 695–1,525, 420–750, and 400–1,600 mel. All the distributions had a variance of 35 mel. No explicit training signals on the categories of the speech sounds were presented.

Neural Map Formation

The model network consisted of a self-organizing population of neurons receiving the speech input described above (Figure 2). Each neuron received as input the activity of the peripheral frequency channels. The connection weights between the frequency channels and the map neurons were adjusted according to two biologically plausible principles: inhibitory competition and Hebbian weight changes (Kohonen, 1982; Sejnowski, 1999; Yuste & Sur, 1999). The activity a of neuron i was determined by

$$a_i = \mathbf{w}_i * \mathbf{p} + \xi.$$

Here, $\mathbf{w}_i$ are the connection weights between the input and neuron i , $\mathbf{p}$ is the peripheral activity pattern presented as input to the model, and ξ is noise from a normal distribution ($M = 0$, $SD = 0.5$).

On each stimulus presentation, the neurons competed for weight change, which is equivalent to having lateral inhibition and only a subset of neurons becoming active. This competition promotes the specialization of model neurons to a limited set of input patterns. Here, this competition was realized through the 10% of the neurons with the highest activity being allowed to change

their weights. The weight changes $\Delta \mathbf{w}$ for neuron i were determined by

$$\Delta \mathbf{w}_i = \varepsilon * a_i * \mathbf{p} - 0.1 * a_i^2 * \mathbf{w}_i,$$

where the first term is the Hebbian weight change. This weight change is determined by the input pattern, and it is proportional to the activity of the winner neuron. When following this learning rule, the winner neuron becomes more likely to respond also to the next presentation of the same input pattern. In other words, the winner neurons become tuned toward the input patterns for which they win the competition against other network neurons. The rate at which the weights changed was controlled by $\varepsilon = 0.05$. The second term prevented excessive growth in the connection weights according to the Oja rule (Oja, 1982). This learning rule was expected to result in a denser representation of frequent than of infrequent input patterns (Guenther & Gjaja, 1996; Kohonen, 1982) and, thus, to create a neuronal representation of speech sounds typical of the environment. The initial connection weights were normally distributed random values ($M = 0$, $SD = 0.05$). Multiple simulation runs showed that the results were not sensitive to the number of neurons in the model. The results shown below are for a population of 100 neurons.

Model Interpretation

The goal of interpreting the neuronal mapping resulting from the self-organization process was, first, to evaluate the perceptual warping produced by the model and, second, to understand which model properties led to warped perception. The perceptual distance between two stimuli was equated with the distance between the activity patterns that they elicited in the neuronal map. The activity patterns can be described as vectors in 100-dimensional space, and, thus, the distance between two patterns can be calculated in the Euclidean metric. With this method, we estimated the perceptual distances between vowel pairs separated by 20 mel in $F1$ or $F2$. The distance measure at each point of the $F1$ – $F2$ space was the average across all vowel pairs centered at the point.

To identify the properties in the model that give rise to warping, the tuning properties of single neurons of the model were evaluated. First, for each neuron, the vowel that elicited maximal activity was identified as the location of maximal tuning, or the best vowel (BV) for that neuron. On the basis of previous results with self-organizing maps (Guenther & Gjaja, 1996; Kohonen, 1982), it was expected that the BVs would concentrate around the frequently occurring input patterns—that is, the prototypes. Second, we calculated the shapes of the tuning curves in $F1$ – $F2$ space and analyzed how this related to modifications in the perceptual distances.

As an additional control condition, all the analyses were performed on networks where the connection weights of a self-organized network were randomly rearranged. Thus, the weights of the randomized control network followed the same distribution as those of the self-organized network. The results were similar for different randomized networks, and, thus, they are presented for one instance

Figure 2. The self-organizing neural network. Connection weights between the model neurons and the peripheral activity pattern serving as input were determined with a Hebbian learning rule.

Figure 3. The statistical distribution of the speech input and maximal tuning of model neurons in the formant ($F1$ – $F2$) space. (A) The input presentation to the model contained four speech categories. Sounds close to the prototypes were presented often, whereas those far from the prototypes were infrequent. (B and C) The speech exposure altered the maximal tuning properties of the network. Each dot in the panels corresponds to one model neuron; its location in the $F1$ – $F2$ space indicates the vowel producing maximal activity in the neuron, and its size is proportional to the magnitude of the connection weights. Tuning in the randomized network was distributed relatively evenly, but in the self-organized network, maximal tuning was concentrated close to the category centers.

only. This allowed us to identify the perceptual warping emerging specifically from the self-organizing process.

RESULTS

Maximal Tuning of Model Neurons

A neural network was allowed to self-organize while speech input was presented to the model as neural activity patterns of the auditory periphery. The model was exposed to a statistical distribution of vowel sounds (Figure 3A), and the resulting neuronal representations of these sounds were evaluated for the neuronal tuning properties and warped perception in $F1$ – $F2$ space. As a control measure, all these measures were also obtained for a randomized network. Each neuron was characterized by its BV—that is, the vowel that activated it maximally. In the randomized network, BVs were evenly distributed across the input space, so that all the areas activated roughly the same number of neurons (Figure 3B). In the network exposed to the speech input, the neurons were more often maximally tuned to prototypical vowels than to other areas in the input space, so that the distribution of maximal tuning followed the distribution of vowel presentation (Figure 3C). More specifically, 64% of the neurons had their BV within a 100-mel radius from a prototype (Figure 4). This pattern of maximal tuning is consistent with previous work (Guenther & Gjaja, 1996) and arises out of the choice of the learning rule (Kohonen, 1982).

Perceptual Distances

To evaluate the perceptual warping following from the speech exposure, perceptual distances between neighboring vowels were calculated for the randomized and the self-organized networks. The two networks differed in both the absolute magnitude and the distribution of the perceptual distances. In the self-organized network, the perceptual distances were, on average, three times longer

and their variation was five times greater than in the randomized network (3.5–8.7 in the self-organized network, 1.5–2.5 in the randomized network). This indicates a general rise in the level of activity to speech sounds in the self-organized network. To allow numerical comparisons between the randomized and the self-organized networks, regardless of the difference in the mean level of activity, the perceptual distance measures were normalized ($M = 0$, $SD = 1$; i.e., they were expressed as z scores). This comparison revealed compression at the prototypical speech sounds resembling the perceptual magnet effect. In the self-organized network, the perceptual distances between neighboring vowels within a 100-mel radius of the prototypes were, on the average, -0.37 (Figure 4). In the randomized network, these distances were -0.08 , indicating a considerable decrease in perceptual distances around the prototypes as a result of self-organization to the speech input. Furthermore, indications of expansion in the perceptual space between the prototypes were also found. Outside a 100-mel radius from the prototypes, the self-organized network had average perceptual distances of 0.10 and the randomized network 0.02. Thus, as a result of the self-organization process, the perceptual distances decreased close to the prototypes and increased for vowels close to the category boundaries, consistent with the perceptual magnet effect and with categorical perception, respectively.

Shape of the Tuning Curves

To examine the relationship between perceptual warping and the tuning properties of single neurons in the model, the $F1$ – $F2$ tuning curves of single neurons were evaluated. The results from a representative set of neurons are shown in Figure 5. Tuning in the model neurons was wide, with a large set of formant values activating each neuron nearly maximally. On average, 33% of the stimuli led to $>80\%$ level of activity in the single neu-

Figure 4. Distance and single-neuron tuning measures in relation to the proximity of a prototype for the self-organized and randomized networks. In the randomized network, the perceptual distances did not depend on the proximity of the prototype. However, in the self-organized network, these distances were much shorter in the vicinity of the prototypes than in regions far from them. In the self-organized network, maximal activity occurred twice as often in the areas close to the prototypes as in the distant areas. In contrast, the randomized network displayed an even distribution of maximal activity in $F1$ – $F2$ space. The highest gradients of the single-neuron tuning curves were located more often far from the prototypes than close to them in the self-organized network. Again, in the randomized network, these gradients were evenly distributed in input space. The lowest gradients of the tuning curves tended to be more often within 100 mel of the prototypes than outside this range in the self-organized network. The opposite was true for the randomized network.

rons. The locations of the highest and the lowest gradients of tuning curve were identified for each neuron in the randomized and self-organized networks (Figure 4). For 64% of the neurons in the self-organized network, the steepest gradient was outside a 100-mel radius of the prototypes, whereas in the randomized network, the steepest slopes were evenly distributed across the $F1$ – $F2$ space. The lowest gradient of the tuning curve fell within a 100-mel radius of a prototype in 57% of the neurons in the self-organized network, as compared with only 26% in the randomized network. Thus, as compared with the randomized network, the self-organized network neurons more often had the highest gradients of their tuning curves far from the prototypes and the lowest gradients close to one

of the prototypes. In conclusion, in the majority of the neurons of the self-organized network, the tuning curves were flat in broad areas centered at the prototypes and steep in the intermediate areas between the prototypes.

How this type of flat tuning leads to long perceptual distances at the category boundary and short distances at the category center is illustrated in Figure 6. Here, the activity of a typical neuron is shown for two pairs of vowel sounds with the same $F1$ but an $F2$ differing by 100 mel. In the first case, the pair of vowels was close to a category center ($F1 = 420$ mel, $F2 = 1,550$ and $1,650$ mel), and in the second, it was midway between two category centers ($F1 = 420$ mel, $F2 = 1,350$ and $1,450$ mel). Close to the category center, both vowels were represented on the flat

Figure 5. The activity level of a representative set of single model neurons to the vowels of the $F1$ – $F2$ space. Each panel represents one model neuron. (A) The activity of the neurons of the randomized network was not organized according to the vowel categories used in the training. (B) In the self-organized network, the best vowels of single neurons were concentrated to the corners of the input space where the prototypes were situated.

top of the tuning curve. Thus, both activated the neuron nearly maximally, and the 100-mel difference in $F2$ led to a very small, 6% change in the level of activity of the neuron (i.e., 0.4 difference). Midway between two category centers, the activity level for the vowels was approximately half of that close to the prototype, but in this region of the stimulus space, the slope of the tuning curve was maximal. Consequently, a 100-mel difference in $F2$ led to a large, 53% change in the activity of the neuron (1.9 difference). Thus, small changes in the stimulation around the category prototypes have very little effect on the level of neuronal activity. In contrast, stimulation changes of the same magnitude at the category boundary lead to large changes in activity. Since the majority (65%) of the neurons behave similarly to the representative neuron examined here, the activity pattern of the whole network mirrors the sensitivity of the representative neurons. That is, when a small perturbation is made in the stimulation, the activity pattern and, therefore, the perceptual distance change more at the category boundaries than at the prototypes.

DISCUSSION

Human perception of speech sounds is characterized by language-dependent warpings of perceptual space (Kuhl,

1991; Liberman et al., 1957). We constructed a computational model that reproduced the two main aspects of these warpings: the expansion at the category boundary and compression around category prototypes. Consistent with human speech learning, speech sounds were presented to the model following a statistical distribution in which the category information was embedded. Furthermore, the auditory input was in the form of neural activity patterns of the auditory periphery. No predetermined phonetic features of the input were extracted or amplified, and no category labels were presented to the model. Instead, the model network was allowed to self-organize following a Hebbian learning rule. The neuronal representation emerging from this speech exposure exhibited both categorical perception and the perceptual magnet effect.

Previous modeling work has attempted to account for perceptual warping in terms of the distribution of neuronal resources, as measured by their maximal activity. The perceptual magnet effect has been explained as arising from the dominance of the neurons representing prototypes in forming the perception of a speech sound (Guenther & Gjaja, 1996). According to this explanation, the category boundaries are underrepresented by the model neurons, and it is unclear how increased discriminability could arise from this underrepresentation. Alternatively, a neu-

Figure 6. Demonstration of how single-neuron tuning properties give rise to perceptual warping. The perceptual distances (A) and the tuning curves of model neurons (B) are plotted as a function of F_2 (while F_1 was kept constant). When two vowel sounds separated by 100 mel were situated close to the prototypical instance that the neuron was maximally tuned to, the activity of the example neuron (black) was nearly equal for the two vowels (Δa_1 , dark gray). However, for another pair of vowels with a 100-mel separation and situated between two prototypes, the difference in the activity of the neuron was almost five times larger (Δa_2 , light gray).

ronal overrepresentation of the category boundaries has been suggested to underlie better discriminability (Bauer et al., 1996; Guenther et al., 1999). In this case, the increased perceptual distances are assumed to follow from an overrepresentation, although this assumption has not been tested. Furthermore, this approach seems counter-intuitive, since it implies that speech processing relies on neuronal specialization to sounds untypical of the speech environment. The present study, in which not only the maximal activity of the model neurons but also the shapes of their tuning curves were considered, offers a framework for explaining perceptual warping in a unified manner. We demonstrate how prototypical instances can be overrepresented in a neural network (as in Guenther & Gjaja, 1996) and, yet, discriminability is increased at category

boundaries without the need to posit neural specialization for these regions. The key to this finding was noting that the neurons take part in encoding all stimuli, not just those that elicit maximal activity in them. Thus, although the neural resources of the full network are available to represent all parts of perceptual space, regions in the vicinity of prototypes are perceptually compressed, and regions at category boundaries are expanded.

Our approach provides a new explanation of warped speech sound perception in terms of neuronal organization. In the present model, the tuning curves of the neurons were wide, and this shape was responsible for both the compressions and the expansions of perceptual space (see Figures 4 and 6). Around the maximal activity coinciding with the category prototypes, the activity varied only slightly in response to small perturbations in the sound stimuli. Consequently, perceptual differences between speech sounds close to the prototype were compressed. At the category boundaries, the level of neuronal activity was lower, but the slope of the tuning curve was steeper. As a result, even small differences between sound stimuli led to large differences in the activity level, and this allowed for better discrimination—that is, perceptual expansion—at the boundary. A similar association between discriminability and the shape of neuronal tuning curves has been demonstrated in the auditory cortex for the representation of sound source location (Stecker, Harrington, & Middlebrooks, 2005; Werner-Reiss & Groh, 2008). Localization is best for sound sources in frontal directions, although the majority of spatially sensitive neurons in the auditory cortex are maximally tuned to sound sources either to the left or to the right. These neurons are widely tuned, so that their activation for sound sources within each hemifield is relatively constant and the largest changes in the level of activity are found for sound sources in the front. A recent modeling study indicated that the wide shape of the tuning curve is, indeed, crucial for the occurrence of maximal activity and best discriminability at different locations in stimulus space (Kim & Bao, 2008). The discrimination ability of a population of neurons was found to depend on the width of the tuning curves. With narrow tuning, best discriminability fell close to the frequency region of maximal tuning, but when wide tuning curves were applied, best discriminability was further away.

The present modeling results seem to be in contrast with those of animal studies that have indicated an enlargement in the tonotopic representation of frequency regions in which the animal shows improvements in perceptual abilities as a result of training (Polley, Steinberg, & Merzenich, 2006; Recanzone, Schreiner, & Merzenich, 1993). Furthermore, when animals are conditioned to respond to a sound stimulus of a specific frequency, the tuning curves of neurons become shifted toward the target frequency, leading to an overrepresentation of the behaviorally relevant stimulus (Bakin & Weinberger, 1990; Blake, Strata, Churchland, & Merzenich, 2002). In the present model, however, the response preferences of the neurons shifted away from the areas where behavioral discriminability was best. The apparent discrepancy between the training-induced experimental effects and those in the present

model can be resolved by considering not only the peak location, but also the width of the neuronal tuning curves. The shifts in the best frequency following conditioning are manifested by neurons sharply tuned to frequency. A narrow tuning curve has its steepest slope near the maximum, and, thus, discriminability is good very close to the stimulus eliciting maximal activity. In contrast, wide tuning, such as that found in the present model, leads to best discriminability far from the preferred stimulus (Kim & Bao, 2008). Thus, due to these differences in tuning width, preferred stimuli coincide with best discrimination in the case of frequency perception but are separated from each other in the case of vowel perception. Analogies should, however, be drawn with caution between speech sound representation and the results of animal conditioning studies. In animal studies, adult subjects are explicitly conditioned to detect a target sound. Speech, in contrast, is learned in infancy, and simply from being immersed in a language environment. Considering the fundamental differences between these learning processes, it also seems plausible that the neuronal mechanisms behind the perceptual abilities arising from them could be different.

Training-induced changes in neuronal representations have also been the subject of human brain-imaging studies. Learning new speech sound categories leads to an increase in the amplitudes of event-related responses to the prototypes of newly learned categories (Tremblay & Kraus, 2002; Tremblay, Kraus, McGee, Ponton, & Otis, 2001). This is consistent with more neurons becoming maximally tuned to the prototypical instances, as suggested by the present model. In contrast, improvements in pitch discrimination resulting from training are accompanied by increased hemodynamical responses (Guenther et al., 1999; Guenther et al., 2004). In other words, the neuronal representation was larger for sound stimuli that were discriminated well. It is, however, unclear how comparable the results obtained by pitch training are to speech processing. The cortical networks engaged in categorizing native speech sounds are different from those involved in learning-induced categorizations (Husain et al., 2006; Luo, Husain, Horwitz, & Poeppel, 2005), indicating that the underlying brain processes might be fundamentally different.

Here, perceptual warping originated from the tuning properties of a neuronal representation self-organized according to sensory input patterns. The network model we used had the tendency to lose its ability to learn new categories after a sufficiently long exposure to one language environment. This might allow the extension of our work to modeling the sensitive period of language learning and the subsequent loss of plasticity. This approach is, however, unable to capture the use of explicit category labels as a cognitive strategy. These are especially important in language learning in adulthood. In recent modeling work, category labels were used in training a network model that replicated the difficulties Japanese subjects have in learning the /r/ and /l/ categories of the English language (Vallabha & McClelland, 2007). In this model, category neurons representing the native speech sounds interfered with the formation of new category representations. Combining this approach with the present one based on self-organization to auditory

peripheral neural activity could be an interesting direction for future research. This would allow the study of the interplay between perceptual warping resulting from passive exposure in infancy and the representations resulting from category labels learned later in life.

AUTHOR NOTE

This study was supported by the Academy of Finland (Project Nos. 111848, 217082, and 217113) and the Emil Aaltonen Foundation. Correspondence concerning this article should be addressed to N. H. Salminen, Department of Biomedical Engineering and Computational Science, Helsinki University of Technology, P.O. Box 3310 (Otakaari 7B), FI-02015 TKK, Helsinki, Finland (e-mail: nelli.salminen@tkk.fi).

REFERENCES

- AALTONEN, O., EEROLA, O., HELLSTRÖM, A., UUSIPAICKA, E., & LANG, A. H. (1997). Perceptual magnet effect in the light of behavioral and psychophysiological data. *Journal of the Acoustical Society of America*, **101**, 1090-1105.
- ANDERSON, J. A., SILVERSTEIN, J. W., RITZ, S. A., & JONES, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, **84**, 413-451.
- BAKIN, J. S., & WEINBERGER, N. M. (1990). Classical conditioning induces CS-specific receptive field plasticity in the auditory cortex of the guinea pig. *Brain Research*, **536**, 271-286.
- BAUER, H.-U., DER, R., & HERRMAN, M. (1996). Controlling the magnification factor of self-organizing feature maps. *Neural Computation*, **8**, 757-771.
- BLAKE, D. T., STRATA, F., CHURCHLAND, A. K., & MERZENICH, M. M. (2002). Neural correlates of instrumental learning in primary auditory cortex. *Proceedings of the National Academy of Sciences*, **99**, 10114-10119.
- DAMPER, R. I., & HARNAD, S. R. (2000). Neural network models of categorical perception. *Perception & Psychophysics*, **62**, 843-867.
- EIMAS, P. D. (1963). The relation between identification and discrimination along speech and non-speech continua. *Language & Speech*, **6**, 206-217.
- FANT, G. (1970). *Acoustic theory of speech production*. The Hague: Mouton.
- FRIEDA, E. M., WALLEY, A. C., FLEGE, J. E., & SLOANE, M. E. (1999). Adults' perception of native and nonnative vowels: Implications for the perceptual magnet effect. *Perception & Psychophysics*, **61**, 561-577.
- GOLDSTONE, R. L., STEYVERS, M., & LARIMER, K. (1996). Categorical perception of novel dimensions. In G. W. Cottrell (Ed.), *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 243-248). Mahwah, NJ: Erlbaum.
- GUENTHER, F. H., & GJAJA, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America*, **100**, 1111-1121.
- GUENTHER, F. H., HUSAIN, F. T., COHEN, M. A., & SHINN-CUNNINGHAM, B. G. (1999). Effects of categorization and discrimination training on auditory perceptual space. *Journal of the Acoustical Society of America*, **106**, 2900-2912.
- GUENTHER, F. H., NIETO-CASTANON, A., GHOSH, S. S., & TOURVILLE, J. A. (2004). Representation of sound categories in auditory cortical maps. *Journal of Speech, Language, & Hearing Research*, **47**, 46-57.
- HARNAD, S., HANSON, S. J., & LUBIN, J. (1991). Categorical perception and the evolution of supervised learning in neural nets. In D. W. Powers & L. Reeker (Eds.), *Working papers of the AAAI Spring Symposium on Machine Learning of Natural Language and Ontology* (pp. 65-74). Kaiserslautern: Deutsches Forschungszentrum für Künstliche Intelligenz.
- HARNAD, S., HANSON, S. J., & LUBIN, J. (1995). Learned categorical perception in neural nets: Implications for symbol grounding. In V. Honavar & L. Uhr (Eds.), *Symbol processors and connectionist network models in artificial intelligence and cognitive modeling: Steps toward principled integration* (pp. 191-206). London: Academic Press.
- HUSAIN, F. T., FROMM, S. J., PURSLEY, R. H., HOSEY, L. A., BRAUN,

- A. R., & HORWITZ, B. (2006). Neural bases of categorization of simple speech and nonspeech sounds. *Human Brain Mapping*, **27**, 636-651.
- IVERSON, P., & KUHL, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Journal of the Acoustical Society of America*, **97**, 553-562.
- IVERSON, P., & KUHL, P. K. (2000). Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from a common mechanism? *Perception & Psychophysics*, **62**, 874-886.
- KIM, H., & BAO, S. (2008). Distributed representation of perceptual categories in the auditory cortex. *Journal of Computational Neuroscience*, **24**, 277-290.
- KOHONEN, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, **43**, 59-69.
- KRUSCHKE, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, **99**, 22-44.
- KUHL, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, **50**, 93-107.
- KUHL, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences*, **97**, 11850-11857.
- KUHL, P. K., ANDRUSKI, J. E., CHISTOVICH, I. A., CHISTOVICH, L. A., KOZHEVNIKOVA, E. V., RYSKINA, V. L., ET AL. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, **277**, 684-686.
- KUHL, P. K., WILLIAMS, K. A., LACERDA, F., STEVENS, K. N., & LINDBLOM, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, **255**, 606-608.
- LIBERMAN, A. M., SAFFORD HARRIS, K., HOFFMAN, H. S., & GRIFFITH, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, **54**, 358-368.
- LUO, H., HUSAIN, F. T., HORWITZ, B., & POEPEL, D. (2005). Discrimination and categorization of speech and non-speech sounds in an MEG delayed-match-to-sample study. *NeuroImage*, **28**, 59-71.
- MACMILLAN, N. A., GOLDBERG, R. F., & BRAIDA, L. D. (1988). Resolution for speech sounds: Basic sensitivity and context memory on vowel and consonant continua. *Journal of the Acoustical Society of America*, **84**, 1262-1280.
- MAYE, J., WERKER, J. F., & GERKEN, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, **82**, B101-B111.
- OJA, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, **15**, 267-273.
- PATTERSON, R. D., ALLERHAND, M. H., & GIGUÈRE, C. (1995). Time-domain modeling of peripheral auditory processing: A modular architecture and software platform. *Journal of the Acoustical Society of America*, **98**, 1890-1894.
- PETERSON, G. E., & BARNEY, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, **24**, 175-184.
- PISONI, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, **13**, 253-260.
- POLLEY, D. B., STEINBERG, E. E., & MERZENICH, M. M. (2006). Perceptual learning directs auditory cortical map reorganization through top-down influences. *Journal of Neuroscience*, **26**, 4970-4982.
- RECANZONE, G. H., SCHREINER, C. E., & MERZENICH, M. M. (1993). Plasticity in the frequency representation of primary auditory cortex following discrimination training in adult owl monkeys. *Journal of Neuroscience*, **13**, 87-103.
- REPP, B. H. (1984). Categorical perception: Issues, methods, findings. *Speech & Language*, **10**, 243-335.
- SAMUEL, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics*, **31**, 307-314.
- SEJNOWSKI, T. J. (1999). The book of Hebb. *Neuron*, **24**, 773-776.
- STECKER, G. C., HARRINGTON, I. A., & MIDDLEBROOKS, J. C. (2005). Location coding by opponent neural populations in the auditory cortex. *PLoS Biology*, **3**, e78.
- TREMBLAY, K. L., & KRAUS, N. (2002). Auditory training induces asymmetrical changes in cortical neural activity. *Journal of Speech, Language, & Hearing Research*, **45**, 564-572.
- TREMBLAY, K. [L.], KRAUS, N., MCGEE, T., PONTON, C., & OTIS, B. (2001). Central auditory plasticity: Changes in the N1-P2 complex after speech-sound training. *Ear & Hearing*, **22**, 79-90.
- VALLABHA, G. K., & MCCLELLAND, J. L. (2007). Success and failure of new speech category learning in adulthood: Consequences of learned Hebbian attractors in topographic maps. *Cognitive, Affective, & Behavioral Neuroscience*, **7**, 53-73.
- 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.
- WERNER-REISS, U., & GROH, J. M. (2008). A rate code for sound azimuth in monkey auditory cortex: Implications for human neuroimaging studies. *Journal of Neuroscience*, **28**, 3747-3758.
- WOOD, C. C. (1976). Discriminability, response bias, and phoneme categories in discrimination of voice onset time. *Journal of the Acoustical Society of America*, **60**, 1381-1389.
- YUSTE, R., & SUR, M. (1999). Development and plasticity of the cerebral cortex: From molecules to maps. *Journal of Neurobiology*, **41**, 1-6.

(Manuscript received September 5, 2008;
revision accepted for publication June 20, 2009.)