From Exemplar Theory to Population Coding and Back. An Ideal Observer Approach.

Laurent Bonnasse-Gahot[†,*] and **Jean-Pierre Nadal**[†,‡]
[†]Centre d'Analyse et de Mathématique Sociales
UMR 8557 CNRS-EHESS
Ecole des Hautes Etudes en Sciences Sociales
54 bd. Raspail, F-75270 Paris Cedex 06
[‡]Laboratoire de Physique Statistique
UMR 8550 CNRS-ENS-Paris 6-Paris 7
Ecole Normale Supérieure
24 rue Lhomond, F-75231 Paris Cedex 05
[*]Corresponding author. Email: lbg@ehess.fr

Abstract

Exploiting the analogy between exemplar models and population coding schemes, we characterize, by means of information theoretic tools, the efficiency of a large but finite population of cells coding for a discrete set of categories (e.g. vowels). The optimal code is shown to typically allocate more cells to class boundaries than to regions further apart. We then discuss the predicted perceptual consequences and review existing exemplar models in the light of our general results.

Keywords: exemplar models, population coding, speech perception, information theory, psychophysics.

1 Introduction

Exemplar models (Hintzman, 1986; Nosofsky, 1986) originally stem from the field of psychology as general models of perception and categorization. They have subsequently been applied to speech perception (Lacerda, 1995; Johnson, 1997) and extended to speech production (Pierrehumbert, 2001, 2003). Although a neuroscientific interpretation has sometimes been mentioned (Lacerda, 1998), such an approach has never been seriously exploited. For instance Kruschke's model, "motivated by a molar-level psychological theory" (Kruschke, 1992), despite a terminology partly borrowed from neuroscience – e.g. *activation, receptive field* –, remains in line with traditional connectionism, as suggested by the use of *node* instead of *neuron* or *cell*.

In this paper, within the framework of speech perception, we propose to take seriously the hypothesis that exemplar models can be given a direct interpretation in term of neural representation. Taking advantage of a recent literature in neuroscience, and making use of standard tools from information theory (see, e.g., Blahut, 1987; Cover and Thomas, 2006), we show not only that this neuroscientific approach is plausible but also that it makes it possible to study in a very general way the optimal neural coding of categories (e.g. vowels), independently of any assumptions about learning or decoding method. This, in turn, will be shown to have consequences for exemplar theory.

This paper is organized as follows. Section 2 sums up the main mathematical result derived in Bonnasse-Gahot and Nadal (2007). In subsection 2.1, we first give a description of the model we use, pointing out its links with both exemplar theory and theoretical neuroscience. We then exhibit in subsection 2.2 the main formula as well as the following predictions. In section 3 we present the perceptual interpretation and consequences of our result. Section 4 considers existing models in the light of our results, and the last section finally gives the concluding remarks.

2 Population coding of categories

2.1 Model description

We assume given a discrete set of categories $\mu = 1, \ldots, M$ (such as phonemes, but our results are applicable to other modalities than speech as well), with probabilities of occurrences $q_\mu \geq 0$,

so that $\sum_\mu q_\mu = 1$. Each category defines a density distribution $P(\mathbf{x}|\mu)$ over the continuous stimulus space (see Pierrehumbert, 2003, p. 119). Along with classical exemplar views, the stimulus space is assumed to be of finite dimensions, those dimensions being those relevant to speech perception (e.g. in the case of vowels, the perceptual dimensions might be the fundamental frequency F0 and the first formants F1, F2, F3). For the sake of clarity, we consider here a one dimensional case : $x \in \mathbb{R}$ (the general case of $\mathbf{x} \in \mathbb{R}^K$ is presented in Bonnasse-Gahot and Nadal, 2007).

Playing the role of N stored exemplars or N (Kruschke's) hidden nodes, we consider a population of N neurons. Each neuron i has an activity specific to a location within the input space, with a mean response given by its *tuning curve* $f_i(x)$, centered around a value x_i (the *preferred stimulus* of cell i, which can be considered as the stored exemplar), and with a width a_i. A typical tuning curve is given by a bell-shaped function, such as

$$f_i(x) = F_i \; exp\left(-\frac{(x-x_i)^2}{2a_i^2}\right)$$

In a standard exemplar model, one would have a uniform value $a_i = a$ of the width. Here, the heterogenity in the widths a_i allows for *local* deformations of the *perceptual space* defined by the output of the neuronal population (we will discuss this point in section 3).
We assume that the responses of the neurons (given x) are not correlated, so that the overall activity $\mathbf{r} = \{r_1, \ldots, r_N\}$ has a factorized probability density function :

$$P(\mathbf{r}|x) = \prod_{i=1}^{N} P_i(r_i|x)$$

with thus

$$\sum_{r_i} P_i(r_i|x) r_i = f_i(x).$$

For the numerical illustration of our results, we will consider that, given an input x, the activity (number of spikes) r_i of the ith neuron is generated according to a Poisson statistics with mean rate $f_i(x)$, that is:

$$P_i(r_i|x) = \frac{\big(f_i(x)\big)^{r_i}}{r_i!} e^{-f_i(x)} \qquad (2.1)$$

This Poisson model is taken here for both its mathematical simplicity and its biological plausibility (see e.g., Tolhurst et al., 1983; Softky and Koch, 1993).

Such a coding is a typical instance of *population coding* (see e.g., Pouget et al., 2000), a strategy widely used in the brain that consists in encoding information by large assemblies of neurons. Two well-known examples are given by the representation of movement direction in the primate motor cortex (Georgopoulos et al., 1986), or by the head-direction cells in rats (Taube et al., 1990). A particularly relevant example here is the inferotemporal cortex in the monkey, which has been shown to be a site for object recognition (see Tanaka, 1996, for a review) and classification. There, population coding is a strategy widely used (e.g. Young and Yamane, 1992; Vogels, 1999), and has already be given an exemplar-based interpretation (Logothetis et al., 1995; Sigala and Logothetis, 2002; Sigala, 2004).
We are interested in quantifying the coding efficiency of such a (neural) representation, and in characterizing the optimal one. Optimality is defined as minimizing the probability of error of an ideal observer during a task of classification. We do not address the question of learning or decoding. We thus do not assume any particular type of learning process nor any decoding method, so that our results remain general and can be applied to any model that shares the same basic assumptions (e.g. Nosofsky, 1986; Kruschke, 1992; Lacerda, 1995; Johnson, 1997).

2.2 Results

Mutual information. In theoretical neuroscience, population coding efficiency has been computed by means of information theoretic tools, in the case of both continuous and discrete stimuli (Seung and Sompolinsky, 1993; Brunel and Nadal, 1998; Kang and Sompolinsky, 2001). We perform a similar analysis in the present context of categorical perception.

A relevant quantity is the *mutual information* that measures the statistical dependency between two variables. Here, we want to compute the mutual information between the set of categories and the neural representation. Maximization of this quantity (which can be the result of learning or adaptation) will have the consequence of minimizing the probability of misclassifying an incoming

stimulus.

The mutual information between the categories μ and the neural activities $\mathbf{r}$ is defined by (Blahut, 1987):

$$I(\mu, \mathbf{r}) = \sum_{\mu=1}^{M} q_\mu \int d^N\mathbf{r}\, P(\mathbf{r}|\mu)\, \log \frac{P(\mathbf{r}|\mu)}{P(\mathbf{r})} \tag{2.2}$$

where $P(\mathbf{r})$ is the probability density function (p.d.f.) of $\mathbf{r}$:

$$P(\mathbf{r}) = \sum_{\mu=1}^{M} q_\mu P(\mathbf{r}|\mu). \tag{2.3}$$

This quantity $I(\mu, \mathbf{r})$ is positive and, by virtue of the data-processing theorem (see e.g. Blahut, 1987), it is upper bounded by the information $I(\mu, x)$ conveyed by the sensory input x about μ. Under mild assumptions one can show that $\lim_{N\to\infty} I(\mu, \mathbf{r}) = I(\mu, x)$.

Large but finite population. In Bonnasse-Gahot and Nadal (2007), we show that for finite but large N the leading correction to this $N \to \infty$ limit is given by:

$$I(\mu, x) - I(\mu, \mathbf{r}) = \frac{1}{2} \int dx\, p(x) \frac{F_{\text{cat}}(x)}{F_{\text{code}}(x)} \tag{2.4}$$

where $p(x)$ is the p.d.f. of the stimulus x: $p(x) = \sum_\mu q_\mu P(x|\mu)$,
and $F_{\text{code}}(x) \geq 0$ is the *Fisher information* characterizing the sensibility of $\mathbf{r}$ with respect to small variations of x (see, e.g., Blahut, 1987):

$$F_{\text{code}}(x) = -\int d^N\mathbf{r}\, P(\mathbf{r}|x)\, \frac{\partial^2 \ln P(\mathbf{r}|x)}{\partial x^2} \tag{2.5}$$

and $F_{\text{cat}}(x) \geq 0$ is the Fisher information characterizing the sensibility of μ with respect to small variations of x:

$$F_{\text{cat}}(x) = -\sum_{\mu=1}^{M} P(\mu|x)\, \frac{\partial^2 \ln P(\mu|x)}{\partial x^2} \tag{2.6}$$

which can also be written as

$$F_{\text{cat}}(x) = \sum_{\mu=1}^{M} \frac{P'(\mu|x)^2}{P(\mu|x)} \tag{2.7}$$

where $P(\mu|x)$ is the probability of having category μ knowing the stimulus x (*ie* the identification function), given, according to Bayes' rule, by

$$P(\mu|x) = \frac{P(x|\mu)q_\mu}{p(x)} \tag{2.8}$$

In (2.7), $P'(\mu|x)$ denotes the derivative of $P(\mu|x)$ with respect to x.

The Fisher information $F_{\text{code}}(x)$ is specific to the coding stage $x \to \mathbf{r}$: it tells how well the neural code can discriminate nearby sensory inputs. The term $F_{\text{cat}}(x) = \sum_{\mu=1}^{M} P'(\mu|x)^2/P(\mu|x)$ is specific of the sensory encoding $\mu \to x$ and thus does not depend on the neural code: it tells how the statistics in the input space are well correlated or not to the categories.

Main qualitative consequences. Typically, an identification function $P(\mu|x)$ has an S-shape, whose slope $|P'(\mu|x)|$ is largest near the boundaries between categories. This entails that the quantity $F_{\text{cat}}(x) = \sum_{\mu=1}^{M} P'(\mu|x)^2/P(\mu|x)$ is greater in these regions. If the code is to be optimized, we therefore expect, as the number N of neurons is limited, Fisher information $F_{\text{code}}(x)$ to be greater between categories than within (see Eq. 2.4). As a consequence, more cells will be devoted to these regions of overlap compared to regions where only one category dominates. This is reminiscent of the Support Vector Machine approach (Cortes and Vapnik, 1995), a technique very popular in machine learning, that identifies exemplars closest to the class boundary as being the most crucial ones for a given classification task. Besides, this result seem to find support in functional imagery and neuro-physiology. First, using functional imagery methods, Guenther et al. (2004) show both for speech and non-speech sounds that category learning entails neural activity in the auditory cortex to be higher in response to stimuli close to the boundary ('non-prototypical' stimuli) than in response to prototypical stimuli (that lie in a more central region of a given category) (see also Guenther and Bohland, 2002, and section 4). Second, in the neuro-physiology of the inferotemporal cortex of the monkey brain, Freedman and colleagues found that category learning leads to a distribution of preferred stimuli mainly peaked at the class boundary: almost half of all the recorded neurons have indeed their preferred stimuli located at the class boundary (Knoblich et al., 2002, Fig. 12).

Numerical Illustration. Figure 1 shows a numerical example involving two categories, whose distributions are shown in Fig. 1.A, and $N = 15$ neurons initially equidistributed. The optimal $\{x_i\}_{i=1}^N$ and $\{a_i\}_{i=1}^N$ are obtained by numerically

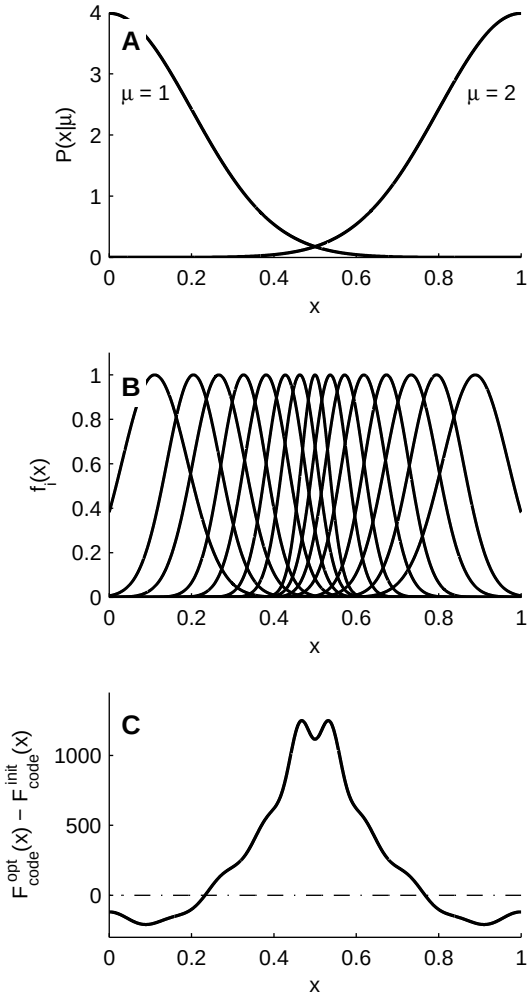

Figure 1: One-dimensional example involving two Gaussian categories. (A) Probability distributions of the two categories. (B) Optimal tuning curves. (C) Difference between the Fisher information $F^{\text{opt}}_{\text{code}}$ for the optimal code, and the Fisher information $F^{\text{init}}_{\text{code}}$ for a equidistributed distribution of preferred stimuli.

maximizing (using simulated annealing) the difference $I(\mu, \mathbf{r}) - I(\mu, x)$ given by equation 2.4. Fig. 1.B shows the resulting tuning curves. As expected, Fisher information $F_{\text{code}}(x)$ (plotted in Fig. 1.C) is the greatest at the boundary between the two categories, and more neurons are allocated in this region compared to regions further apart. Note also that the tuning curves are sharper near the boundary, *ie* the width a_i of the corresponding cells is narrower than the width of cells away from the boundary. One could see that this width plays the role of the inverse of the 'attentional weight' found in classical exemplar-based models (Nosofsky, 1986; Kruschke, 1992). In other words, more local 'attention' is devoted to the class boundary, further sensitizing the neuronal population to this region. Note that this is a collective effect, coming from having both an heterogeneous set of widths (each cell i having its own a_i) and a specific distribution of preferred stimuli.

3 Towards an explanation of categorical perception

As previously stated, if the code is optimized, the Fisher information $F_{\text{code}}(x)$ is greater at the boundary between categories than within (Eq. 2.4). The Fisher information $F_{\text{code}}(x)$ is linked to the discriminability d' (from signal detection theory, and commonly used in psychophysics; see e.g. Green and Swets, 1988) of two stimuli x and $x + \delta x$ according to (Seung and Sompolinsky, 1993) :

$$d' = |\delta x| \sqrt{F_{\text{code}}(x)} \tag{3.9}$$

Thus, a perceptual consequence of maximizing mutual information between neural responses and categories, under the constraint of a fixed number of cells, is that discriminability d' will be greater at the boundary between categories, a phenomenon traditionally called *categorical perception* (Harnad, 1987).

Categorical perception was first described within the field of speech perception as an innate process specific to human speech that implied high discriminability between items from different (phonemic) categories and zero discriminability within a category (Liberman et al., 1957). This strong version of categorical perception was subsequently undermined: not only this phenomenon can be acquired (Abramson and Lisker, 1970; Francis and Nusbaum, 2002) but it is also not specific to speech (Goldstone, 1994; Livingston et al., 1998; Özgen and Davies, 2002) nor to human (Kuhl and Padden, 1983; Kluender et al., 1998). Moreover, the all-or-none effect on discriminability was never found experimentally: within-category differences are discriminable. Our result fits well into this framework, for it can apply to any modality, might be induced by learning, and does not assume anything specific to human. Besides, the discriminability within a category is not zero.

Another way to present the effects induced by the adaptation of the neural configuration is in terms of compression/expansion of the perceptual space defined by the output of the neuronal population. If discriminability is higher (respectively lower) after learning than before, then the perceptual space can be seen as expanded (resp. contracted). Category learning might imply different outcomes. For example, using visual stimuli, Goldstone (1994) found *acquired distinctiveness* at the class boundary (increased between-categories differences), whereas Livingston et al. (1998) found *acquired similarity* (increased within-category similarity). In the case of the learning of new phonetic categories, Francis and Nusbaum (2002) reported both compression and expansion of the perceptual space. Whether category learning induce within-category compression and/or between-category expansion might depend on the initial ability of the neuronal population. In the numerical illustration given by figure 1, we see that, compared to the equidistributed initial configuration of preferred stimuli, there is *acquired distinctiveness* at the boundary and *acquired similarity* within categories. Such a warping of the perceptual space is related to a phenomenon found in speech perception literature called the *perceptual magnet effect* (Kuhl, 1991), stating that discriminability is lower around prototypical stimuli than around non-prototypical ones, even if the corresponding stimuli belong to the same category. These prototypicality effects (as well as frequency effects) will be more extensively studied in a forthcoming paper (Bonnasse-Gahot and Nadal, in preparation).

An important aspect of our model is that we do not need category labels so as to find categorical effects, which might shed light on one of the most disputed issues on categorical perception. This question, that finds its roots in the Whorfian hypothesis stating that our language shapes our vision of the world (Kay and Kempton, 1984), paradoxically concerns the very basis of this phenomenon: is categorical perception really perceptual ? Views are divided. Some argue that this phenomenon is not perceptual after all but only results from the use of verbal labels (Roberson and Davidoff, 2000), whereas others maintain that categorization does alter perception (Goldstone et al., 2001; Notman et al., 2005). Our view follows the latter. Our result indeed gives an optimal bound on discriminability, based on a purely sensory level, and thus gives credit to a perceptual account for the two phenomena described above, namely categorical perception and perceptual magnet effect. Note, however, that we do not claim that other processes, such as top-down influences, memory effects, or labeling, might not intervene in discriminability judgments.

To conclude, our result indicates that categorical perception is not a mere by-product of category learning but serves a function, that is to minimize classification errors. This gives a quantitative theoretical support to the Native Language Neural Commitment posited by Kuhl (2004) stating that language experience induces neuronal modifications that aim at enhancing the features relevant for native language but entail difficulties in the learning of a foreign language (see, e.g., Kuhl et al., 1992; Iverson et al., 2003).

4 Comparison with existing models

In this section we want to reconsider existing models in the light of our results. Two main kinds of models, designed to account for categorical perception and/or perceptual magnet effect, are concerned : exemplar-based models (Lacerda, 1998; Goldstone et al., 1996) and neural maps (Bauer et al., 1996; Guenther and Gjaja, 1996; Guenther and Bohland, 2002). Note that all these models share the same basic assumptions concerning the architecture: a perceptual map is covered by 'cells' or 'exemplars' centered around a preferred stimuli with the responsiveness of their receptive field decreasing as the incoming stimulus moves away from the preferred stimulus. The models differ by how this map is decoded (e.g. with a specific additional layer), and/or by the learning algorithm used to build the map. In our case, we have characterized the coding efficiency of a map (see equation 2.4), independently of any learning process or decoding method. Taking as a reference the expected properties of an optimal code, we can thus compare our predictions with the empirical results obtained using specific algorithms.

As we have seen in section 2.2, a direct consequence of equation 2.4 is that category centers are represented by fewer cells (neurons,

exemplars) than category boundaries, which in turn explain *why* perceptual phenomena such as categorical perception or the perceptual magnet effect might arise (section 3). This is in line with results obtained by Goldstone et al. (1996), Bauer et al. (1996), Guenther and Bohland (2002) but in disagreement with models proposed by Guenther and Gjaja (1996) or Lacerda (1995, 1998). Let us review these two latter models in more depth.

Interestingly, Guenther and Gjaja (1996) model is the only one from the above list for which the learning algorithm is not specifically meant to solve a categorization task. Following traditional self-organizing map approach, the learning mechanism leads to a situation where the distribution of preferred stimuli of the coding cells follows the distribution of stimuli. This actually corresponds to a situation of density estimation (representing the distribution of x, instead of coding for the categories), and should lead to higher discriminability where the distribution of stimuli peaks, *ie* at the center of the categories, contrary to a situation of categorical perception. Guenther et al. (1999) have experimentally shown that the same distribution of stimuli can lead to opposite perceptual outcomes, depending on the training task (discrimination vs categorization). Guenther therefore proposed a different version so as to take into account the need for categorization. This time, the subsequent model allocates more cells to the boundary (Guenther and Bohland, 2002; Guenther et al., 2004), in agreement with our qualitative predictions.

Let us now turn to Lacerda's model (Lacerda, 1995, 1998). It basically assumes that every encountered exemplar is stored, leading again to more cells around the modes of the distribution of stimuli, contrary to our result. In order to explain the perceptual phenomena discussed above, a discrimination measure involving the category label of the exemplars is introduced. This runs counter to the fact that labels might not be used in a discrimination task, especially in the case of the magnet effect for which it is assumed that all items belong to the same category. For instance, in Iverson and Kuhl (1995), subjects were conditioned to view all items as belonging to the same category, which might prevent them from using category labels, hence working on a more perceptual basis.

More generally, our result goes counter to the most common working hypothesis in computational exemplar models that all encountered exemplars are stored, leading to a higher density of cells at the center of a category. It also sheds light on the "head-filling up problem" (Kruschke, 1992; Johnson, 1997; Pierrehumbert, 2001) that the more traditional approach has to face, as the memory is of finite size. It indeed shows that some exemplars are more informative about categories than others, and that much compression can thereby be achieved at the center of a category.

5 Conclusion and Future Works

To sum-up, we first gave a neural interpretation of exemplar models in terms of population coding of categories, which makes it possible to address the question of optimal coding with information theoretic tools that have been shown to be relevant in computational neuroscience. We find that the mutual information between the activity of a neuronal population and a set of discrete categories is simply given by the average over the input space of the ratio of the Fisher information of the categories over the Fisher information of the neuronal population. As seen in section 2.2, the category related Fisher information is typically the greatest at the boundary between categories. As neural resources are limited, this entails that, if there is adaptation, the Fisher information of the neuronal population should also be the greatest between categories in order to minimize the probability of misclassifying an incoming stimulus. As this Fisher information is directly related to the discriminability, category learning implies better cross-category discrimination than within-category discrimination, which gives an explanation of categorical perception (see section 3).

As we have seen, we make no assumption on the learning mechanism nor on the decoding method. This makes it possible to evaluate existing models (section 4) and more generally to establish a groundwork for future work. A possible direction for future research will now consist in studying these mechanisms in more depth.

Let us first get into the question of decoding. Several methods such as population vector or maximum likelihood have already been proposed

in theoretical neuroscience, so that this point might not be a technical issue. More interestingly, one can ask whether decoding is necessary. Empirical research has indeed shown that categorical information as well as fine phonetic details within-categories are used by the listener (Miller, 1994; McMurray et al., 2002). In that spirit, our results might give clues on how two different levels of representation, one continuous and concrete (stimulus) and one discrete and abstract (categories), can be combined within the same neural representation. On the one hand, thanks to population coding, information about the stimulus and its detailed properties might be retrieved, but on the other hand, because of the neural modifications induced by category learning (language experience), this neural representation also contains information about the category the stimulus belongs to.

Concerning the question of learning, cross-language studies have demonstrated that infants of 6 months of age are already tuned to their native language, as in the case of the perception of vowels (Kuhl et al., 1992), well before they can talk or have acquired a lexicon. Moreover, many experiments have recently shown that both adult, infant, and animal listeners are able to extract distributional information from the input signal (Saffran et al., 1996; Guenther et al., 1999; Lotto, 2000). Maye et al. (2002) showed not only that 6 and 8-month old infants are sensitive to the statistical distribution of the speech sounds they hear in their environment but also that this in turn influences their discrimination ability. These results call for the study of *unsupervised* learning that would aim at attaining the ideal state described by our result. Such a study will have to face a paradoxical issue given by our result. More resources have indeed to be allocated to category boundaries that are typically regions where exemplars are rarely encountered. As a consequence, one might ask how information about the distribution of stimuli might be extracted from regions of low-frequency 'traffic'. Beyond technical issues, this question is particularly interesting from a developmental perspective.

Acknowledgments

This work is part of a project "Acqlang" supported by the French National Research Agency (ANR-05-BLAN-0065-01). LBG acknowledges a fellowship from the DGA. JPN is a CNRS member.

We thank Sharon Peperkamp and Janet Pierrehumbert for introducing us to this topic and for valuable discussions. LBG is also grateful to Emmanuel Dupoux for numerous and stimulating discussions. We also thank three anonymous referees for useful comments and suggestions.

References

Abramson, A. and Lisker, L. (1970). Discriminability along the voicing continuum: Cross-language tests. In *Proceedings of the Sixth International Congress of Phonetic Sciences*. Prague: Academia.

Bauer, H.-U., Der, R., and Herrmann, M. (1996). Controlling the magnification factor of self-organizing feature maps. *Neural Computation*, 8(4):757–771.

Blahut, R. E. (1987). *Principles and practice of information theory*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Bonnasse-Gahot, L. and Nadal, J.-P. (2007). Neural coding of categories. (submitted).

Bonnasse-Gahot, L. and Nadal, J.-P. (in preparation).

Brunel, N. and Nadal, J.-P. (1998). Mutual information, fisher information, and population coding. *Neural Computation*, 10:1731–1757.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Cover, T. and Thomas, J. (2006). *Elements of Information Theory*. Wiley & Sons, New York. Second Edition.

Francis, A. and Nusbaum, H. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, 28(2):349–366.

Georgopoulos, A., Schwartz, A., and Kettner, R. (1986). Neuronal population coding of movement direction. *Science*, 233:1416–1419.

Goldstone, R. (1994). Influences of categorization on perceptual discrimination. *J Exp Psychol Gen.*, 123(2):178–200.

Goldstone, R., Lippa, Y., and Shiffrin, R. (2001). Altering object representations through category learning. *Cognition*, 78:27–43.

Goldstone, R., Steyvers, M., and Larimer, K. (1996). Categorical perception of novel dimensions. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*, pages 243–248, Hillsdale, New Jersey. Lawrence Erlbaum Associates.

Green, D. and Swets, J. (1988). *Signal detection theory and psychophysics, reprint edition*. Los Altos, CA: Peninsula Publishing.

Guenther, F. and Bohland, J. (2002). Learning sound categories: A neural model and supporting experiments. *Acoustical Science and Technology*, 23(4):213–221.

Guenther, F. and Gjaja, M. (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., Husain, F., Cohen, M., and Shinn-Cunningham, B. (1999). Effects of categorization and discrimination training on auditory perceptual space. *Journal of the Acoustical Society of America*, 106:2900–2912.

Guenther, F., Nieto-Castanon, A., Ghosh, S., and Tourville, J. (2004). Representation of sound categories in auditory cortical maps. *Journal of Speech, Language and Hearing Research*, 47(1):46–57.

Harnad, S., editor (1987). *Categorical Perception: The Groundwork of Cognition*. New York: Cambridge University Press.

Hintzman, D. (1986). "schema abstraction" in a multiple-trace memory model. *Psychological Review*, 93(4):411–428.

Iverson, P. and Kuhl, P. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Acoustical Society of America Journal*, 97:553–562.

Iverson, P., Kuhl, P., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Ketterman, A., and Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, 87:B47–57.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In Johnson and Mullenix, editors, *Talker Variability in Speech Processing*, pages 145–165. San Diego: Academic Press.

Kang, K. and Sompolinsky, H. (2001). Mutual information of population codes and distance measures in probability space. *Physical Review Letters*, 86(21):4958–4961.

Kay, P. and Kempton, W. (1984). What is the sapir-whorf hypothesis. *American Anthropologist, New Series*, 86(1):65–79.

Kluender, K., Lotto, A., Holt, L., and Bloedel, S. (1998). Role of experience for language-specific functional mappings of vowel sounds. *J. Acoust. Soc. Am.*, 104(6):3568–3582.

Knoblich, U., Freedman, D., and Riesenhuber, M. (2002). Categorization in it and pfc: Model and experiments. *AI Memo 2002-007. Cambridge, MA: MIT AI Laboratory*.

Kruschke, J. (1992). Alcove : An exemplar-based connectionist model of category learning. *Psychological Review*, 99(1):22–44.

Kuhl, P. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Percept Psychophys*, 50(2):93–107.

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

Kuhl, P. and Padden, D. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *J. Acoust. Soc. Am.*, 73(3):1003–1010.

Kuhl, P., Williams, K., Lacerda, F., Stevens, K., and Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255:606–608.

Lacerda, F. (1995). The perceptual-magnet effect: an emergent consequence of exemplar-based phonetic memory. In Elenius, K. and Branderyd, P., editors, *XIIIth International Congress of Phonetic Sciences*, volume 2, pages 140–147, Stockholm.

Lacerda, F. (1998). Distributed memory representations generate the perceptual-magnet effect. *Journal of the Acoustical Society of America*.

Liberman, A., Harris, K., Hoffman, H., and Griffith, B. (1957). The discrimination of speech

sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54:358–369.

Livingston, K., Andrews, J., and Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 24(3):732–753.

Logothetis, N., J.Pauls, and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, 5(5):552–563.

Lotto, A. (2000). Language acquisition as complex category formation. *Phonetica*, 57:189–196.

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

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

Miller, J. (1994). On the internal structure of phonetic categories: a progress report. *Cognition*, 50:271–285.

Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology*, 115(1):39–57.

Notman, L., Sowden, P., and Özgen, E. (2005). The nature of learned categorical perception effects: a psychophysical approach. *Cognition*, 95:B1–B14.

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

Pierrehumbert, J. (2003). Phonetic diversity, statistical learning and acquisition of phonology. *Language and Speech*, 46(2-3):115–154.

Pouget, A., Zernel, R., and Dayan, P. (2000). Information processing with population codes. *Nature Review Neuroscience*, 1(2):125–132.

Roberson, D. and Davidoff, J. (2000). The categorical perception of colors and facial expressions: The effect of verbal interference. *Memory & Cognition*, 28:325–340.

Saffran, J., Aslin, R., and Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, 274:1926–1928.

Seung, H. S. and Sompolinsky, H. (1993). Simple models for reading neuronal population codes. *Proceedings of the National Academy of Science*, 90:10749–10753.

Sigala, N. (2004). Visual categorization and the inferior temporal cortex. *Behavioural Brain Research*, 149:1–7.

Sigala, N. and Logothetis, N. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, 415:318–320.

Softky, W. and Koch, C. (1993). The highly irregular firing of cortical cells is inconsistent with temporal integration of random epsps. *The Journal of Neuroscience*, 12(1):334–350.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.*, 19:109–139.

Taube, J., Muller, R., and J.B. Ranck, J. (1990). Head-direction cells recorded from the postsuiculum in freely moving rats. i. description and quantitative analysis. *The Journal of Neuroscience*, 10(2):420–435.

Tolhurst, D., Movshon, J., and Dean, A. (1983). The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Research*, 23:775–785.

Vogels, R. (1999). Categorization of complex visual images by rhesus monkeys. part 2: single-cells study. *European Journal of Neuroscience*, 11:1239–1255.

Young, M. and Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science*, 256:1327–1330.

Özgen, E. and Davies, I. (2002). Acquisition of categorical color perception: A perceptual learning approach to the linguistic relativity hypothesis. *Journal of Experimental Psychology: General*, 131(4):477–493.