Available online at www.sciencedirect.com
ScienceDirect
Cognition 103 (2007) 80–106
COGNITION
www.elsevier.com/locate/COGNIT

Learning phonetic categories by tracking movements ☆

Bruno Gauthier [a,*], Rushen Shi [a], Yi Xu [b]

[a] *Département de psychologie, Université du Québec à Montréal, C.P. 8888, Succursale Centre-Ville, Montréal, Que., Canada H3C 3P8*
[b] *Department of Phonetics and Linguistics, University College London, London, UK*

Received 8 June 2005; revised 25 February 2006; accepted 1 March 2006

Abstract

We explore in this study how infants may derive phonetic categories from adult input that are highly variable. Neural networks in the form of self-organizing maps (SOMs; Kohonen, 1989, 1995) were used to simulate unsupervised learning of Mandarin tones. In Simulation 1, we trained the SOMs with syllable-sized continuous F_0 contours, produced by multiple speakers in connected speech, and with the corresponding velocity profiles ($D1$). No attempt was made to reduce the large amount of variability in the input or to add to the input any abstract features such as height and slope of the F_0 contours. In the testing phase, reasonably high categorization rate was achieved with F_0 profiles, but $D1$ profiles yielded almost perfect categorization of the four tones. Close inspection of the learned prototypical $D1$ profile clusters revealed that they had effectively eliminated surface variability and directly reflected articulatory movements toward the underlying targets of the four tones as proposed by Xu and Wang (2001). Additional simulations indicated that a further learning step was possible through which $D1$ prototypes with one-to-one correspondence to the tones were derived from the prototype clusters learned in Simulation 1. Implications of these findings for theories of language acquisition, speech perception and speech production are discussed.

☆ This manuscript was accepted under the editorship of Jacques Mehler.
* Corresponding author.
E-mail address: gauthier.bruno@courrier.uqam.ca (B. Gauthier).

0010-0277/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2006.03.002

Keywords: Category formation; Infant speech perception; Language acquisition; Unsupervised learning; Self-organizing maps; Target approximation; Lexical tone; Contextual tonal variation; Theories of speech production and perception

1. Introduction

The task of learning the sounds of the native language is daunting. Infants do not receive explicit language instructions, nor are they able to make inquiries about the structure that they are learning. They must discover the phonetic categories of their native language from the speech input of the surrounding speakers. The task is further complicated by the fact that they do not know how many categories to discover along any particular input dimension. To make things worse, the input they receive is highly variable. That is, there is lack of invariant acoustic manifestation of phonetic categories. A classical example was discussed by Peterson and Barney (1952), who demonstrated how between-category formant frequencies show great overlap in the American English vowel space when produced by multiple speakers. Many sources of variability have been studied since, including coarticulation, spoken rhythm and dialectal variations. Phonetic categories other than vowels have also been shown to be subject to different sources of variability. Liberman (1970) pointed out how English /b, d, g/ produced in the same vowel context show continuously changing second formant transition slope patterns with identifiable category boundaries separating the three sounds. Subsequent perception studies have shown good agreement between these acoustic characteristics and the perception of stop consonants (Menon, Rao, & Thosar, 1974; Ohde, 1988). However, these general slope patterns can change drastically in certain vowel contexts. For example, the second formant slope for /d/ is positive before /i/ and negative before /u/ (Delattre, Liberman, & Cooper, 1955). The variable second transition slope patterns for /d/ overlap to some degree with the slope patterns of /b, g/ in certain vowel contexts.

The variability problem is just as severe when it comes to lexical tones. In many languages, words are distinguished from one another not only by consonants and vowels, but also by pitch patterns that occur during the voiced sound (Yip, 2002). In Mandarin, for example, the syllable /ma/ can mean "mother", "hemp", "horse" or "to scold" depending on whether its pitch pattern is high-level (High tone), rising (Rising tone), low-dipping (Low tone) or falling (Falling tone). The primary acoustic correlate of tones is F_0, i.e., the fundamental frequency of voice (Abramson, 1962; Chao, 1933; Howie, 1976). Although other phonetic/prosodic cues have been suggested to contribute to the perception of tones (e.g., duration, amplitude (Whalen & Xu, 1992); voice quality for languages with lexical phonation types (Andruski & Ratliff, 2000; Maddieson & Hess, 1986)), F_0 has been consistently shown to be the dominant cue in adult tone perception (e.g., Klein, Zatorre, Milner, & Zhao, 2001; Whalen & Xu, 1992). Fig. 1(a) shows the (time-normalized) F_0 contours of five tokens and their means of the four Mandarin tones produced in citation form by a

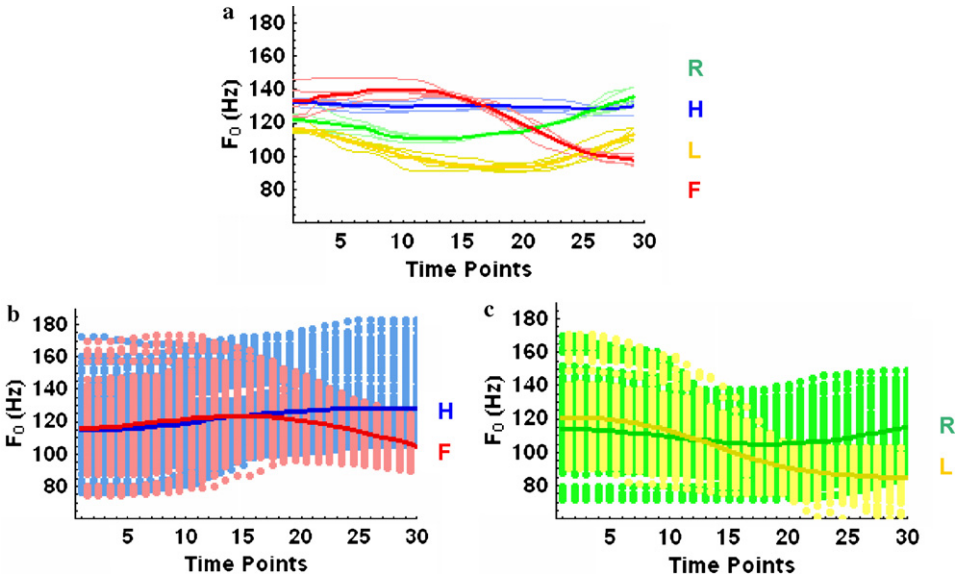

Fig. 1. Tones produced (a) in citation form by one speaker and (b, c) in connected speech by three speakers. Thick lines correspond to means while pale background to the distribution of High (blue), Rise (green), Low (yellow) and Fall (red) (data from Xu (1997)).

male speaker (data from Xu (1997)). As can be seen, when produced in isolation by a single speaker, the tones are well separated even when time-normalized. They become much less separated, however, when spoken in connected speech and when uttered by different speakers. Figs. 1(b) and (c) show the means and distributions of the same four tones spoken in connected utterances by three male speakers (Xu, 1997). The extensive overlap between the tones comes from two major sources.[1] The first is the difference in the pitch range of individual speakers and the second is the variability introduced by tonal context in connected speech (Shen, 1990; Xu, 1994, 1997). Similar variability has been found in other tone languages such as Thai, Vietnamese and Yoruba (Han & Kim, 1974; Gandour, Potisuk, & Dechongkit, 1994; Laniran & Clements, 2003).

While tonal variability is similar to segmental variability in nature, tones typically involve a single primary acoustic dimension, namely, F_0. This contrasts with the multiple acoustic dimensions such as formants or spectral peaks required for characterizing vowels and consonants. The variability problem with tones is therefore at least limited to a single dimension, which makes them ideal for testing hypotheses that involve detailed mechanisms of phonetic acquisition. In the present study, we will therefore use lexical tone as a probing tool to find a breaking point for

[1] There are many other sources of variability in tonal realization, as discussed in detail in Xu (2001, 2005). However, most of the other sources of variability are kept constant in the data shown in Fig. 1.

understanding how infants could develop phonetic categories from adult speech input that is highly variable.

Although there has been much research on tone perception by adults, most studies have only investigated the perception of tones produced in isolation (e.g. Abramson, 1978; Gandour, 1983; Karlgren, 1962; Liang, 1963; Massaro, Cohen, & Tseng, 1985; Moore & Jongman, 1997; Shen & Lin, 1991; Whalen & Xu, 1992). Studies on speech input to infants have shown, however, that about 90% of parental speech to infants is multi-word utterances (e.g., Brent & Siskind, 2001; Shi, Morgan, & Allopenna, 1998; Weijer, 1998). In addition, infants generally hear multiple speakers in their daily life. Strikingly, young learners in tone languages seem to have their preliminary tonal categories in place before the first year of life. In fact, their earliest comprehension vocabulary and discriminative abilities already demonstrate certain knowledge of tonal categories. Perception studies by Harrison (2000) and Mattock (2004) indicate that tone-learning infants attend more closely to phonemic tones than to non-significant pitch variations at six months, and that their tonal perception is influenced by the phonemic systems of their ambient language at nine months. Moreover, given that these studies used stimuli with some degree of variability, the results also show that infants can handle certain variability in tonal perception. Although there is still no direct evidence about the age at which infants can deal with between speaker and contextual variability in tonal perception, research on infant speech perception has shown that perceptual normalization of segmental variability in vowels and consonants has already happened in infants before the onset of speech production (e.g., Kuhl, 1976, 1979, 1983; Kuhl & Miller, 1982; Jusczyk, Pisoni, & Mullenix, 1992). Since speech input to infants consists mainly of multi-word utterances by multiple speakers, tone learning must also involve processes that can not only effectively resolve the speaker and contextual variability, but also discover the number of tonal categories as well as the invariant characteristics of each category. But the question is, of course, how can infants do it?

To understand how it is possible for infants to discover tonal categories despite substantial variability and inter-tonal overlap, it is necessary to understand the nature of the variability. Of the two sources of variability mentioned earlier, the nature of cross-speaker difference in pitch range is quite transparent. The length and thickness of the vocal folds vary extensively across gender, age and individuals. The mean pitch therefore differs from speaker to speaker (Zemlin, 1988). The nature of contextual variability is more complex, but much has been learned about it in recent research (e.g., Gandour et al., 1994; Xu, 1997, 1999). For example, much of the variability in both Mandarin and Thai is induced by the preceding tone, i.e., due to carry-over effect, although the following tone exerts anticipatory effect in some contexts. For example, if a High tone in Mandarin is preceded by a Low tone, it will be realized with a rising contour in the earlier part of the tone. Nevertheless, for all the four tones, regardless of what the preceding tone is, the F_0 contours of the syllable associated with the tone all gradually converge over time to an asymptote that is characteristic of the underlying tone: high-level for High, low-rising for Rise, low-level for Low and high-falling for Fall (Xu, 1997, 1999). Based on these findings, Xu and Wang (2001) proposed a theoretical model of contextual tonal realization: the

Target Approximation (hereafter TA) model to account for the contextual variability of tones in production.

In the TA model, surface F_0 patterns are characterized as asymptotic movements toward underlying pitch targets defined as simple linear functions. These targets can be either static or dynamic. Static targets are specified by relative pitch height (e.g., [high], [low]) and dynamic targets by both the relative height and velocity of the pitch movement (e.g., [rise], [fall]). The pitch target therefore is the articulatory goal associated with the lexical tone. The articulation of both static and dynamic targets is subject to the physical constraints of: (1) the maximum speed of pitch change due to the properties of the laryngeal muscles and mechanical characteristics of the larynx (Sundberg, 1979; Xu & Sun, 2002) and (2) the coordination of the larynx and other articulators (Kelso, 1984; Kelso, Saltzman, & Tuller, 1986; Xu & Wang, 2001). The maximum speed of pitch change has been demonstrated to be rather slow so that pitch movements are frequently made as fast as possible during regular speech (Xu & Sun, 2002) but still leading to recurrent undershoot (Xu, 1999). This means that surface F_0 contours consist mostly of movements toward one tonal target or another. The constraint of coordination of the larynx with other articulators has been argued to result in full synchrony between laryngeal and supra-laryngeal movements, so that the F_0 movement toward each tonal target is made only within the syllable that the tone is associated with (Xu & Wang, 2001).[2]

Based on the TA model, an intriguing prediction can be made. That is, despite the extensive contextual tonal variability as well as the tonal overlap due to pitch range differences across speakers, it is possible to infer the underlying pitch targets from the manners of F_0 movements even without context and speaker information, assuming that syllabic segmentation has been done. This can be achieved by taking the first derivative of F_0 (henceforth $D1$), which is the velocity of F_0 movement. $D1$ reflects the characteristics of F_0 movement toward the underlying pitch target. Moreover, as there exist pitch range variations due to speaker differences and intonational factors (Xu, 2005), the transformation of F_0 to $D1$ automatically eliminates most of these pitch range differences. For example, suppose the surface F_0 contour of a tone is represented by a polynomial of the form

$$y = a + bx + cx^2 + \cdots + mx^n. \tag{1}$$

Taking the first derivative of (1) reduces it to

$$y' = 0 + b + 2cx + \cdots + nmx^{n-1}. \tag{2}$$

The transformation turns a, the y-intercept of the polynomial, uniformly to 0, thus normalizing the initial F_0 height, which contains information both about the speaker and about the preceding tone. While both kinds of information are useful, they are not directly relevant to the tone to be recognized. Although Xu (1994) has shown the usefulness of contextual information for the recognition of severely distorted tones

[2] The TA model has been used to explain various tonal and segmental data in both tonal and non-tonal languages (e.g., Chen & Xu, 2006; Xu & Liu, in press; Xu & Xu, 2005). It has also been tested in speech synthesis (Sun, 2002; Prom-on, Xu, & Thipakorn, 2006).

when they are produced in a prosodically weak position (the second syllable in a trisyllabic word) in Mandarin, it is not known whether context information is always indispensable. A major question we pose in the present study is therefore: Is the information about movements toward underlying tonal targets as represented by $D1$ sufficient for the categorization of the four Mandarin tones produced in connected speech by multiple speakers? A positive answer would point to a powerful tool that listeners may have in their possession for disentangling the vast amount of variability without the help of contextual information and complex normalization schemes. More importantly, for infants who are born into a Mandarin-speaking community with presumably no pre-endowed tone categories, a positive answer would mean that they are actually able to use this powerful tool for deriving from the adult input the underlying pitch targets associated with the lexical tones even if they have not yet developed effective strategies for taking advantages of the contextual information and for normalizing speaker differences.

In speech perception research, it is often assumed that some kind of feature extraction needs to take place in order to recognize a sound as belonging to one category or another. Similarly, in tone perception studies, proposed solutions typically try to single out an acoustic feature such as height or slope of F_0 contours corresponding to each tone (Abramson, 1978; Gandour, 1983; Massaro et al., 1985; Shen & Lin, 1991; Wang, 1967). Thus for both segmental and tonal perception, there is a popular assumption that some kind of preprocessing is done to single out certain abstract features in order to perform categorization. Note that, however, preprocessing, even if it does occur, would be just as difficult as the categorization task itself, because it would still need to first resolve the variability problem. If, instead, phonetic categories could be discovered by directly tracking continuous movements in the acoustic signal, the need for feature extraction would be greatly reduced. Thus, another major question we will ask in the study is this: Is it possible to derive phonetic categories directly from continuous signal input without extraction of abstract features?

It has been known for a long time that as they grow older, infants become less sensitive to non-native phonemic contrasts which are non-phonemic in their native language (Werker & Tees, 1984; Werker & Lalonde, 1988), as well as to sounds of foreign languages that are similar to those in their first language (Best, 1993). Such a reduction in sensitivity becomes even more extreme in adults (Best, McRoberts, & Goodell, 2001). At the same time, however, it has been demonstrated that even adults have not totally lost their ability to discriminate sounds that bear little resemblance to any sound in their native language (Best et al., 2001) or to establish a new category division in sounds that belong to the same category in their native language (Bradlow, Pisoni, Yamada, & Tohkura, 1997). These observations have lead to perception theories such as the Functional Reorganization Model (Werker & Tees, 1984), the Native Language Magnet Theory (NLM: Kuhl, 1991) and the Perceptual Assimilation Model (PAM: Best, 1995). Little is known, however, about how these behavior patterns are linked to the core mechanisms of the learning process itself. The third question we will ask in the present study is therefore: Is there a possible link between the decline in sensitivity to within-category differences and the core mechanisms of learning phonetic categories?

2. Methodology

To test the possibility that $D1$ can be used in the perception and the learning of tones in Mandarin Chinese, we use a self-organizing neural network known as the self-organizing map (SOM; Kohonen, 1989, 1995). SOM is a statistical pattern recognition device using unsupervised learning methods for discovering the structure of high dimensional data. It is a particular case of neural map for which the basic idea comes from the discovery of topographically organized projections from the periphery to cortical areas in the brain (Kohonen, 1982). Information encoding by topographic maps has been observed in many regions of the brain, including some areas of the auditory cortex (Crottaz-Herbette & Ragot, 2000; Pantev et al., 1994; Seldon, 1985; Wessinger, Buonocore, Kussmaul, & Mangun, 1997). Because of its visualization properties, the SOM is useful for exploring the internal properties of the data as well as for modeling the place coding of sound properties in brain topographic maps. The model is also consistent with an inductive account of speech perception development. Recent research has shown that infants are sensitive to the statistical distributional properties of speech sounds in the input (Maye, Werker, & Gerken, 2002). Similarly, the SOM decodes systematic statistical patterns found in input distributions. The structural and functional properties of the SOM, combined with the simplicity of its algorithm, thus provide an attractive method for revealing invariants in the speech signal in general, and for modeling the learning of tone categories by naïve learners in particular. The detailed algorithm of the model is presented in Appendix A.

Our simulations attempt to verify how continuous F_0 and $D1$ perform as input to the SOM and how $D1$ function as a normalizing parameter for both between speaker and contextual variability. Two simulations were performed to test the effectiveness of category formation through unsupervised learning, with either syllable-sized F_0 profiles obtained from a natural data set (Xu, 1997) or the syllable-sized velocity profiles ($D1$) derived from those F_0 profiles as input. Simulation 1 used a large receptive mapping area (with many units) for training and testing. Simulation 2 used a much smaller mapping area but with prototypes developed in Simulation 1 as training input and the same F_0 and $D1$ profiles as in Simulation 1 as testing input.

2.1. Simulation 1

2.1.1. Input coding

The input corpus contains 1800 exemplars of the four Mandarin tones produced in connected utterances by three adult male speakers (data from Xu (1997)). Each stimulus corresponds to the first or second syllable of disyllabic 'mama' produced in the middle of a carrier sentence which had either high or low pre-target F_0 offset and post-target F_0 onset. Each input token is a 30 data point vector composed of equal-distanced discrete values taken from a syllable-sized time-normalized F_0 curve (for the exact F_0 extraction procedure, see Xu (1997)). The data are first transformed from Hertz to the Bark scale according to

$$F_{0\ \mathrm{bk}} = 7 \cdot \mathrm{Log}[F_{0\ \mathrm{hz}}/650 + ([1 + (F_{0\ \mathrm{hz}}/650)^2]^{1/2})]. \quad (3)$$

The Bark scale is a frequency scale corresponding to human auditory perception. It is logarithmic at high frequencies but linear in low frequencies. Thus the transformation has no other impact than rescaling the pitch patterns in a way that facilitates future comparisons between F_0 and $D1$ mappings. In the input corpus, F_0 values range from 50 to 180 Hertz and $D1$ from -13 to 9. By using barks (0.5 to 2) F_0 and $D1$ are more comparable in scale. The $D1$ profiles are generated according to

$$D1 = 0.5(F_{0\ \mathrm{hz}}(t+1) - F_{0\ \mathrm{hz}}(t-1)), \quad (4)$$

which yields input vectors of 28 dimensions representing the discrete first derivatives of F_0 patterns.

2.1.2. Learning phase

During the adaptation process, the SOM implements a regression algorithm to map a continuous input distribution $P(x)$, $x_i \in \mathrm{X}$, onto a discrete output space—a 10×10 map consisting of 100 units. The training corpus contains 900 stimuli, which are randomly presented to the network for 100 times. Each time the neighborhoods on the map are shifted to better fit the data.

2.1.3. Testing phase

During the recall task, new exemplars are used to verify the network's capability to generalize to novel data. The trained network assigns each input pattern, from a new set of 900 tokens, to a single unit using the transmission rule described in Appendix A. The testing corpus, which contains as much variability as the learning one, is presented in an orderly fashion. The procedure involves presenting, in order, all exemplars of High, i.e., tone 1, (240 tokens), Rise—tone 2 (240), Low—tone 3 (180) and Fall—tone 4 (240) tones. The Low category contains fewer exemplars because a tonal variant of this tone due to a sandhi rule has been removed from the training and testing sets (Low tone becomes Rise when followed by another Low tone in Mandarin. Cf. Xu, 2001).[3]

2.1.4. Output coding

The trained networks are squared arrays of 10×10 processing units, each one being tuned to a particular subset of input patterns. During the testing phase, the number of input patterns projected onto each unit is indexed into a global firing frequency matrix. Units which fire at least once during recall according to the transmission rule (see Appendix A) are considered as operable units. If a unit never fires during recall, it is considered non-operable. The number of activations of each unit for each class of input patterns is also indexed into four tone firing

[3] Tonal variation due to sandhi is a problem beyond the scope of the current project, because as far as surface acoustics is concerned, the sandhi-derived Rising tone resembles the underlying Rising tone so closely that listeners do not hear them as different tones (Peng, 2000; Wang & Li, 1967). This tonal sandhi alternation for Tone 3 involves context-sensitive rule learning at a morpho-phonological level. It is a learning process entirely different from the type of learning that we are testing in the current study.

frequency submatrices, the sum of which corresponds to the global matrix. The proportion between tone firing frequencies and the global firing frequency of a unit yields the tone probability for this unit, i.e. its probability to be activated by each class given the testing corpus. Tone probabilities give rise to the distinction between categorized and ambiguous units. Units that have a tone probability equal to or above 68% are considered as categorized and are labeled with that particular tone. Units without such a majority class are considered as ambiguous. Such units respond to multiple tones, but none of the tones is dominant. This criterion is decided based on the central limit theorem. That is, plus and minus one standard deviation from the mean includes 68% of the responses to a particular tone.

2.2. Measures

2.2.1. Quantitative criteria

Performance and reliability measures are first used to assess the global properties of the maps. The first performance measure is categorical error. It corresponds to the proportion of the network which responds to more than one class, i.e., the number of ambiguous units on the total number of operable units. The second performance measure is classification error. It corresponds to the probability of the network to respond ambiguously during recall. The test tokens which land on ambiguous units are considered errors. The classification error is thus the number of error tokens divided by the total number of input tokens in the testing corpus. For example, if half of the testing corpus activates ambiguous units during recall, the classification error would correspond to $450/900 = 0.5$. The performance measures help to quantify the clustering properties of the trained maps and they reflect the amount of category information carried by the input distributions.

The most common measures of reliability assessment of the trained SOM are the quantization and the topology errors. The quantization error, which evaluates the precision of the mapping, is given by

$$e_q = 1/n\left(\sum \|x_i - r_v\|\right), \tag{5}$$

where x_i corresponds to the input pattern and r_v to the best matching unit (BMU) for that pattern. The equation gives the sum of the distance between each input pattern and its BMU divided by the total number of stimuli. It thus yields the average distance between input vectors and their BMU's receptive field center. The second reliability measure is the topology error, which evaluates the topographical organization of the map. It is given by

$$e_t = 1/n\left(\sum d(x_i)\right), \tag{6}$$

where $d = 0$ if the first and second BMUs for a given input are next to each other and $d = 1$ if they are not. The equation thus indexes 1 every time topology is not respected and divides the final amount by the total number of stimuli. Reliability

measures are usually used to ensure the good functioning of the SOM so as to validate the conclusions inferred from other results about a data set. In the present study, they also act as a window on F_0 and $D1$ input distributional properties.

The preceding measures are simple scalar summaries for describing the clustering and distributional properties of the maps. A more detailed analysis of groups of units is useful for observing within and between-category map properties. In this regard, the between-category assessment of each condition can be expressed in terms of confusion patterns between each tone and will be presented in the form of confusion matrices. Finally, the rate of success measured for each tonal category is obtained by the sum of input tokens which activate corresponding labeled units divided by the amount of input tokens belonging to this category. For example, if all High input patterns activate High categorized units, the rate of success for High is $240/240 = 1.0$.

2.2.2. Visualization of the maps

Projection techniques are used for graphically revealing the distributional and neighboring structure of the trained maps. Traditional ways to visualize the state of the SOM include the Sammon's mapping and other derived techniques. For example, the u-matrix is a regular grid of neurons between which the relative distance is represented in tones of grey; the lighter the color, the closer neurons are to each other. Another popular technique is the data histogram, which shows how many stimuli belong to a cluster defined by each neuron. Visualization can also be done by projecting the weight vectors into a color space in which similar units are assigned similar colors (e.g., Varfis, 1993; Kaski, Venna, & Kohonen, 1999). In this study, the coloring of the maps is done by representing tone categories with four distinct colors produced with the CMYK color system. The High tone is represented by blue, specified by a mix of cyan and magenta in the vector $[1,1,0,0]$; Rise $= [1,0,1,0]$; Low $= [0,0,1,0]$ and Fall $= [0,1,1,0]$. Each map unit is thus described by a four-dimensional vector where the last element (black) remains null and where the other elements are specified in terms of the firing probabilities for each tone. If each category is well separated in the data, the "color map" should be divided into regions by classes. When a unit responds to more than one tone, the colors associated with each tone are mixed to yield 'impure' colors. Fig. 2(a) shows the legend for interpreting the color maps.

A more conventional way to visualize the final state of the network is the phoneme map (Kohonen, 1989, 1995). This technique assigns each processing unit a label corresponding to the majority class of that unit. While the color map produces a clear display of whether regions of clusters are formed, the phoneme map reveals more precisely the confusion areas. Fig. 2(b) shows an idealized phoneme map of 4 units.

Finally, the internal maps can be used to infer important characteristics and subtle details of a data set. A internal map is a graphical display of the connections which link the input space to the output space. More precisely, it is a projection of the receptive field center of the output units onto the input space. For example, the internal map of a network with n output units connected to

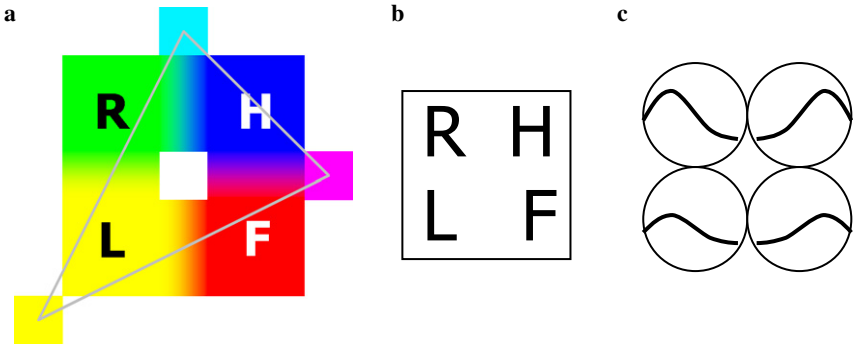

Fig. 2. (a) Color legend of the color map using the CMYK color system (see text for details); (b) idealized 2×2 phoneme map; (c) idealized 2×2 internal map.

a two-dimensional input space shows n data points on a plan, each of which corresponds to an output unit's receptive field center. In this study the input space as well as each output unit's receptive field center are 28- or 30-dimensional. The internal maps of F_0 and $D1$ thus project the prototype vector of each unit in a frequency or velocity by time space for visualizing the patterns developed during training. The prototype vectors can be observed separately as in Fig. 2(c), where four neighboring units display similar temporal patterns. The F_0 and $D1$ internal maps are presented later in a different style, overlaying the prototypes of each class on a single graph for a more direct comparison of groups of map units. Before engaging in such a detailed analysis, we consider whether or not $D1$ is a strong normalization procedure.

3. Results and discussion

In this section, the results and their interpretations are presented with respect to different aspects of the trained maps. The first part describes and compares the maps' global properties for both F_0 and $D1$ conditions. The second part focuses on the comparisons between groups of map units formed by the F_0 and $D1$ training corpus. Finally, the description of individual map units is presented mainly for the $D1$ condition.

3.1. Global map and input distributional properties

The global maps show directly on the trained network, rather than in the input space, whether topologically ordered categories are present in the data.

3.1.1. Performance results

Table 1 shows categorization errors (column 2) and classification errors (column 3) for F_0 and $D1$. The categorization error is larger for F_0 than for $D1$

Table 1
Categorization and classification errors of the performance measures for F_0 and $D1$ conditions

	Performance measures	
	Categorization error	Classification error
F_0	0.20	0.22
$D1$	0.03	0.03

($0.20 > 0.03$), indicating that a majority of the $D1$ map units are category-specific while a larger portion of the F_0 map contains ambiguous units. Fig. 3(a) shows how ambiguous units form widely spread confusion areas on the F_0 map. In contrast, in Fig. 3(b) the $D1$ map shows a cleaner division of regions by classes, thus better representing the categories. These results suggest that the input distribution of $D1$ contains more categorical information than does the F_0 distribution.

Unlike the categorical error measure, which assesses the units' responses, the classification error measure represents the percentage of new tokens that are ambiguously classified. The results (column 3 in Table 1) indicate a higher proportion of tokens being ambiguously classified in the F_0 map (0.22) than in the $D1$ map. The performance with $D1$ (0.03) is again nearly error-free, as in the case of the categorical error measure. In terms of input properties, this means that the density of the F_0 input distribution is greater than that of the $D1$ distribution in overlapping regions, i.e., more tokens are present in the overlapping region in the F_0 input space.

3.1.2. Reliability results

The quantization and topology error for each map are given in Table 2. The quantization error is higher for $D1$ than for F_0 ($0.24 > 0.12$), indicating that the average distance between $D1$ input patterns and their BMU is twice that of F_0's. These results, combined with the performance results, show that minimization of error accomplished by training does not necessarily give a better recognition rate.

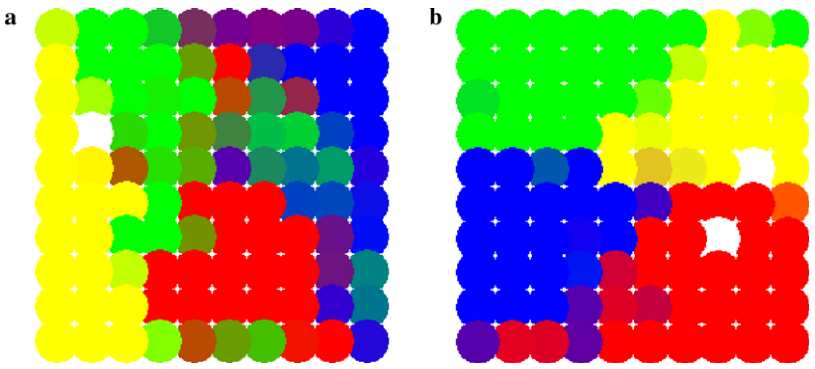

Fig. 3. Color maps of (a) F_0 and (b) $D1$ after training.

Table 2
Quantization and topology errors of the reliability measures for F_0 and $D1$ conditions

	Reliability measures	
	Quantization error	Topology error
F_0	0.12	0.009
$D1$	0.24	0.005

In the present context, category separation is more important than error minimization, and the result suggests that quantization error being too low might actually impede the ability of the network to detect between-category differences. This is because a lower quantization error corresponds to a higher number of units each having a small receptive field, which in turn might indicate a more compact global density function of the input distribution. The more compact a distribution is, the harder it may be to distinguish between neighboring data points which in fact could belong to different classes. According to this argument, $D1$ better represents tonal categories because it stretches the input space in such a way as to enhance the between-category contrasts.

The topology conserving property of the SOM indicates whether or not the input distributions under study possess intrinsic organization, i.e., neighborhood structure. The F_0 topology error is higher than that of $D1$ ($0.009 > 0.005$), suggesting that the F_0 data set contains more discontinuities and that velocity profiles form more coherent clustering of tonal categories. $D1$ thus seems better suited than F_0 for topographical representation. The topology error can also indicate the proportion of the input tokens corresponding to a particular category which might switch class due to only small variations. In this sense, the F_0 system is more sensitive to noise than $D1$.

3.1.3. Summary of global results

The analyses in this section first showed that more distinct clusters are formed on the $D1$ map than on the F_0 map and that the probability for these clusters to be activated during recall was much higher for $D1$ than for F_0. The quantization and topology errors indicated that $D1$ was a more reliable cue for tone recognition and better suited for topographical representation of the input.

3.2. Groups of map units and input manifold properties

In this section we examine more specific aspects of the maps and of the data sets.

3.2.1. Between-category ambiguity

The phoneme maps in Fig. 4 show the categorized (single labels) and ambiguous units (multiple labels) of F_0 and $D1$. The F_0 map contains more ambiguous units, which also show greater diversity among the categories they confuse. For the 20 ambiguous units of F_0, four units confuse tones H and R, three confuse tones H and F, two confuse tone R and L, four confuse tones R and F, and seven units confuse more than two categories. In contrast, the $D1$ map contains fewer

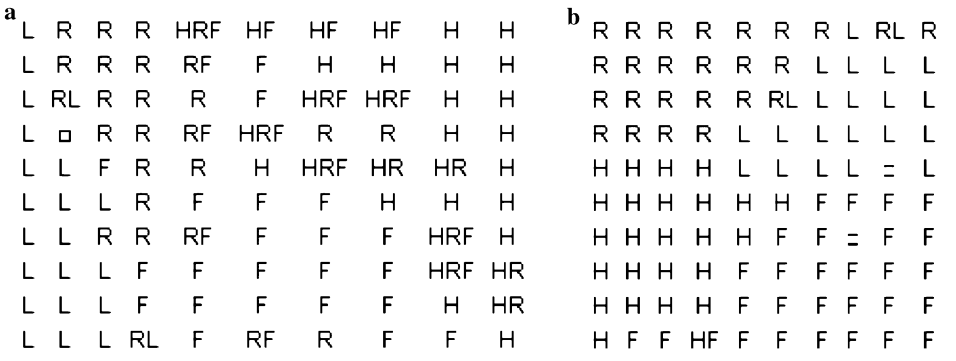

Fig. 4. Phoneme maps of (a) F_0 and (b) $D1$: categorized (single label) and ambiguous (multiple label) units. Squared label correspond to non-operable units.

confused units. Specifically, two units confuse tones R and L and a single unit confuses tones H and F. Overlap is thus present between each tonal category in F_0 while the $D1$ map shows a more systematic confusion pattern. Another observation is that $D1$ overlapping regions are situated between the two classes they separate, while this is not always the case with F_0. For example, the HF cluster in the middle top region of the F_0 map appears between the High and Rise classes and the HR cluster (low right region) is between High and Fall regions. The $D1$ map thus seems more adequate for characterizing the boundaries between tonal categories.

For revealing more subtle confusion patterns between each tone, Table 3 presents a confusion matrix where the rows correspond to the speakers' intended targets and the columns to the majority class of the units responding to the input patterns. The last column shows the number of tokens for each category which activate ambiguous units. For example, in the F_0 condition, of the 240 intended High tone targets, 162 are classified correctly to the High units, 6 misclassified to Rise units and 72 landed

Table 3
Confusion matrix for F_0 and $D1$ conditions

	Confusion matrix				
	High	Rise	Low	Fall	Ambiguous
F_0					
High	162	6	0	0	72
Rise	**11**	157	**8**	**6**	58
Low	0	0	172	0	8
Fall	**11**	**12**	1	160	56
D1					
High	219	1	2	**8**	10
Rise	2	226	3	0	9
Low	0	0	172	1	7
Fall	**9**	0	1	224	6

on ambiguous units. As can be seen, the total number of misclassification is lower for $D1$ (27) than for F_0 (55). Also, the darker elements of the matrices, which show higher probability misclassification patterns, are in greater number for the F_0 map, in which tone R gets mostly misclassified to tones H, L and F and tone F to tones H and R. The $D1$ condition appears to show only a single misclassification pattern between tones H and F.

3.2.2. Within-category rate of success

In the $D1$ confusion matrix, the number of tokens assigned to the corresponding majority class is overall higher than in F_0, as shown by each matrix diagonal element. This agrees with other results gathered so far. The Low tone in F_0, however, behaves differently, which leads us to consider each category in detail. The within-category rate of success for each tonal category is shown in Table 4. The rate of success corresponds to the number of times categorized units respond to a corresponding intended target divided by the total number of tokens of this category in the testing corpus. For example, of the 240 High tones presented to the network in the F_0 condition, 162 are projected onto a High unit, yielding a rate of success of $162/240 = 0.68$ for the High category.

These results show that in the F_0 condition, High, Rise and Fall tones share a similar rate of success of about 66%, while the Low tone enjoys a success rate of 96%. Learning from F_0 information thus allows the network to only recover the Low tone with a high level of accuracy. In contrast, the results from the $D1$ condition indicate that every category shares a similar high rate of success which varies between 91 and 96%. Together with the confusion pattern results, the rate of success brings further evidence that $D1$ better represents tonal categories.

3.3. Simulation 2: Modeling the abstraction of categories after clustering formation

The results of Simulation 1 suggest that the $D1$ profiles of the four Mandarin tones provide sufficient information for a naïve system with no pre-existent tone categories to develop distinct cluster regions for the four tonal categories with well-defined boundaries in between. To answer question 3 raised in the *Introduction*, we test in a new simulation whether the learning system is able to further abstract from the learned maps four distinct categories. Based on the assumption that the new process simulated is neurologically linked to that of the first

Table 4
Within-category rate of success for F_0 and $D1$ conditions

	Rate of success	
	F_0	$D1$
High	0.68	0.91
Rise	0.65	0.94
Low	0.96	0.96
Fall	0.67	0.93

simulation, the neural map consists of the same number of units as the number of clusters learned in Simulation 1.

3.3.1. Methodology

Instead of the 10×10 array used in Simulation 1, the neural map is now a two-dimensional array of 2×2 units. The training corpus now consists of 100 input profiles of F_0 or $D1$, each corresponding to a unit prototype vector developed in Simulation 1. During training, the learning parameter is kept the same as in Simulation 1, but the neighborhood function has been adjusted to better fit the size of the new map, reducing its values by a factor of 10 (i.e. 100–>1 becomes 10–>0.1). This is reasonable given that smaller radius is more appropriate for a four-unit network, as opposed to the 100-unit network in Simulation 1. The testing phase, as well as the output coding, are identical as in Simulation 1 and the same measures are applied for assessing the trained maps.

3.3.2. Results

Table 5 shows performance measures for F_0 and $D1$ in Simulation 2. The categorical error (i.e., the percentage of ambiguous units) and classification error (percentage of tokens landing on ambiguous units) are high for F_0, but the performance of $D1$ is much more successful. Although the perfect performance of $D1$ may be partially related to an artifact of the performance measures, we conducted the same analysis for Simulation 2 as for Simulation 1 to maintain consistency across simulations. Fig. 5 shows phoneme maps of F_0 and $D1$ after training. As can be seen, the categorization with the F_0 input is poor, with all four units responding confusingly to multiple categories. The $D1$ input, on the other hand, resulted in four units representing four distinct tones, suggesting successful abstraction of four tonal categories from the well-delineated clusters and neighborhoods developed during the previous training using 100 units.

4. General discussion

At the outset of the study we raised three questions: (a) Is it possible for a perceptual system to derive phonetic categories directly from continuous signal input without extraction of abstract features? (b) Is the information about movement toward underlying tonal targets as represented by $D1$ sufficient for the categorization of the four Mandarin tones produced in connected speech by multiple speakers?

Table 5
Performance measures for F_0 and $D1$ conditions in Simulation 2

	Categorization error	Classification error
F_0	1.00	1.00
$D1$	0.00	0.00

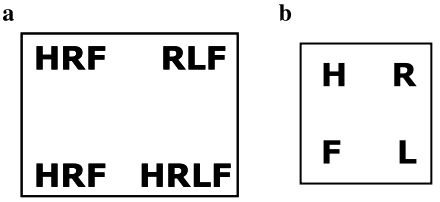

Fig. 5. Phoneme maps of (a) F_0 and (b) $D1$ after training in Simulation 2. The phoneme map of F_0 shows four ambiguous units while the phoneme map of $D1$ shows four categorized units.

(c) Is there a possible link between the decline in sensitivity to within-category differences and the core mechanisms of learning phonetic categories?

To answer the first question, we used the self-organizing map method that is biologically plausible because it is based on the discovery that there are topographically organized projections from the periphery to cortical areas in the brain (Kohonen, 1982). The biological plausibility of our methodology is further enhanced by the naturalness of the input that we used. In all the simulations, time-varying continuous trajectories were used directly as input, with no further extraction of more abstract features. The only preprocessing involved are: (a) the extraction of continuous F_0 trajectories from the acoustic signal, (b) conversion of F_0 trajectories to continuous velocity profiles, and (c) division of the trajectories into syllable-sized chunks. Both (a) and (b) are highly plausible in human perception, given what is known about pitch perception and velocity processing and representation in human brains (Gandour, 1983, 2000; Seldon, 1985). The division of continuous profiles into syllable-sized chunks, i.e., (c), seems reasonable on a number of considerations. The syllable appears intuitively salient, as evidenced by the facts that many, including the Chinese writing systems represent speech directly at the level of the syllable (Chao, 1968; DeFrancis, 1984), and that linguistic theories typically treat the syllable as a level of representation (e.g. Chomsky & Halle, 1968; Prince & Smolensky, 1993). More importantly, experimental research on speech perception has shown that the syllable is the perceptual units in very young infants (Bertoncini & Mehler, 1981; Jusczyk & Derrah, 1987; Bijeljac-Babic, Bertoncini, & Mehler, 1993). There has also been accumulating evidence for the syllable as a critical unit in speech production as summarized in recent theories of the syllable (Fujimura, 2000; Krakow, 1999; MacNeilage, 1998; Xu & Liu, in press). It is therefore plausible to assume that some kind of division of the acoustic signal into syllable-sized chunks occurs in the brain during learning in infants and processing in adults.

As shown by the results of Simulation 1, even the lowest performance, obtained with F_0 as input, achieved 80% correct categorization. Thus the answer to the first question is positive: It is indeed possible for a biologically plausible perceptual system to learn at least one type of sound categories directly from continuous signal input without extraction of abstract features. The implication is that phonological features such as those proposed by Jacobson, Fant, and Halle (1952) and Chomsky and Halle (1968), while seemingly appealing to us as scientific observers, may not be the elements actually processed by the brain.

What is more likely to be processed, as suggested by the even better performance of $D1$ in Simulation 1 (97% correct categorization), is the actual movements themselves. And such high performance has provided a clear positive answer to the second question, namely, the information about movement toward underlying tonal targets as represented by $D1$ is sufficient for the categorization of the four Mandarin tones produced in connected speech by multiple speakers. To better understand why the $D1$ is so much more effective than F_0, we plotted in Fig. 6 the internal maps, which show the prototypical F_0 and $D1$ profiles developed during training in Simulation 1 for each tone category. Two general patterns can be observed. First, the F_0 profiles show much larger within-category vertical spread than $D1$ profiles, and the spread is especially wide near the syllable onset. Second, the F_0 profiles show much less distinct movement patterns than $D1$ profiles. In fact, with the only exception of Low, F_0 profiles of each tone move in both overall directions: up and down. The $D1$ profiles, in contrast, display high consistency in terms of the overall direction of movement. And they differ within each tonal category mostly in magnitude of the movement.

The consistent $D1$ profiles seem to directly reflect the nature of the F_0 movements as characterized by the Target Approximation model (Xu & Wang, 2001) and by the velocity profiles of movements proposed by Nelson (1983). Considering the static tones, most of the High profiles increase their speed from 0 toward positive values, reaching peak velocity around the center of the syllable and finally slowing down toward the initial speed of 0 near the end of the syllable. The Low profiles show almost mirror images of the High profiles. Such unimodal velocity profiles fit the definition of a simple movement given by Nelson (1983), i.e., one that starts from one

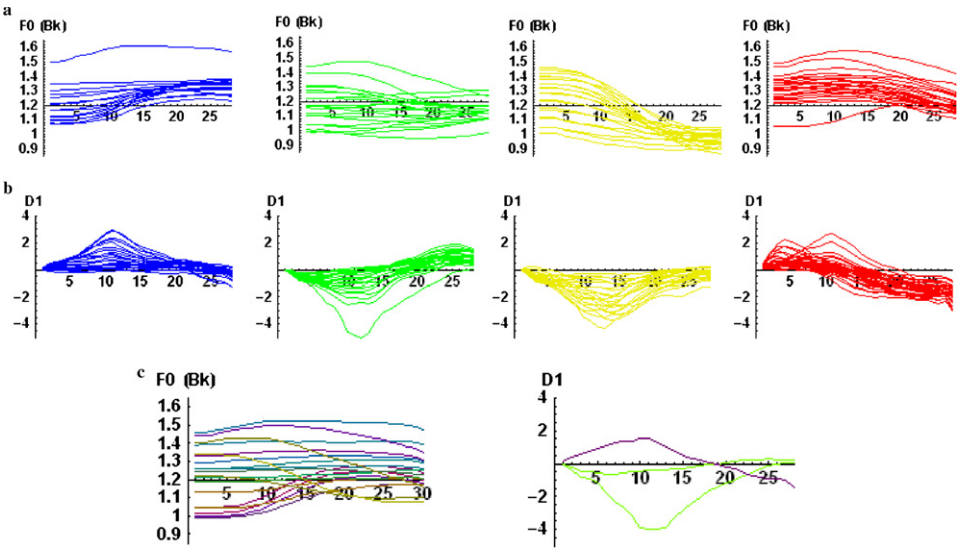

Fig. 6. Internal maps of the four tones in (a) F_0 and (b) $D1$ categorized units and (c) F_0 and $D1$ ambiguous units.

position and stops at another. A voluntary movement such as reaching satisfies this definition. It follows then that the movements involved in the High and Low tones are those toward a static F_0 height. The positive velocity profiles during High correspond to movements toward an above-average pitch height, and the negative velocity profiles during Low correspond to movements toward a below-average pitch height.

The $D1$ profiles of the dynamic tones present a different picture. Like the static tones, the $D1$ profiles of Rise and Fall both increase their speed from 0 at syllable onset toward a negative/positive value. But instead of continuing with the initial direction, the $D1$ profiles reverse their directions, cross the zero speed line and continue until a high velocity is reached near the end of the syllable. In other words, the Rise/Fall velocity profiles indicate rapid initial F_0 movement toward a relatively low/high F_0, followed by another movement in the opposite direction toward the zero line, thus indicating a movement toward an initial static height per Nelson's (1983) definition. But the movements afterwards no longer fit Nelson's definition. Rather, the fact that $D1$ reaches a high (positive or negative) value near the end of the syllable suggests that the high velocity itself is the final goal of Rise and Fall. In other words, the targets of these tones are dynamic, i.e., with a simple linear function as their goal, as postulated in the Target Approximation model (Xu & Wang, 2001).

Based on the above understanding, the prototypes developed in Simulation 1 actually contain three apparently inappropriate ones. One in Rise (with a very low valley) that should belong to the Low tone, and two in the Fall tone (with a very low valley) that should belong to the High tone. Indeed, the further categorized $D1$ profiles developed in Simulation 2 (Fig. 7) seem to have fully eliminated those deviant prototypes.

The direct characterization of articulatory movement toward underlying tonal targets is not the only benefit of $D1$ profile as input to a learning system. It also has the benefit, as explained in the *Introduction*, of immediately removing most of the individual differences in terms of their idiosyncratic pitch ranges as well as much of the influence of the preceding tone. Variability due to both of these sources can be

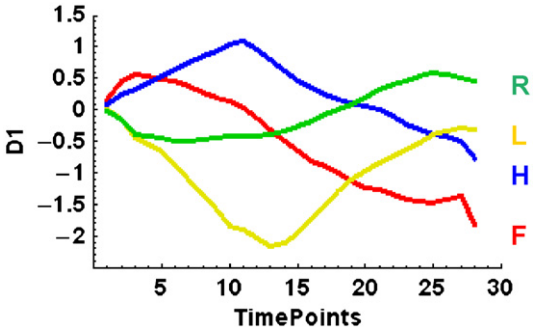

Fig. 7. Velocity profiles for the four units after Simulation 2.

clearly seen in the large vertical distribution of initial F_0 in the upper row of Fig. 6. Such variability is virtually absent in the $D1$ profiles in the lower row of the figure. As explained in the *Introduction*, the differentiation process eliminates from a function the term specifying its y-intercept, thus removing most of the speaker-related variability and much of the context-related variability.

The results of Simulation 1 suggest that with sufficient number of receptive units in a neural network, variability in $D1$ profile can be handled by developing topographical clusters with well-defined borders, each corresponding to a tonal category. While the formation of clusters in Simulation 1 may seem to support the hypothesis of the exemplar theory that linguistic categories can be represented by probabilistic exemplar clouds (e.g., Pierrehumbert, 2002), the results of Simulation 2 suggest that exemplar clouds do not need to be always maintained. That is, a further learning step can take place in which the learned profile clusters can be further reduced to even more ideal prototypes with one-to-one correspondence to the tonal categories. Such direct representations achieved through a two-stage learning process suggest the possibility of drastically reducing the number of neurons needed to represent phonetic categories in the later stage.

An implication of such more economical representation is that it could naturally lead to the behavior patterns described by speech perception theories such as the Functional Reorganization Model (Werker & Tees, 1984), the Native Language Magnet Theory (Kuhl, 1991) and the Perceptual Assimilation Model (Best, 1995), namely, as children mature, their sensitivity to sounds of foreign languages is reduced, unless those sounds bear little resemblance to any sound in the children's native language. Thus, the results of Simulation 2 suggest a positive answer to the third question of the present study, i.e., there is indeed a possible link between the decline in sensitivity to within-category differences and the core mechanisms of learning phonetic categories. It is conceivable that the increased efficiency in a native language as a child matures is related to the precedence given to the mapping of incoming speech sounds to the prototypes formed in the second stage of learning. Such precedence would bias the incoming foreign sounds toward the closest prototype whenever possible. But the precedence can be also softened if a new round of cluster formation is made to happen, as during extensive perceptual training (Bradlow et al., 1997) or during any focused second language learning.

Perhaps the biggest surprise to us was the success of the tonal category formation without direct assistance of any contextual information. This is because we have seen in previous research that the availability of contextual information to the listener is critical for recognizing tones severely distorted by tonal contexts (Xu, 1994). We note, however, a major difference between the input data used in the present study (from Xu, 1997) and those used in Xu (1994). That is, the target tones in the latter were produced in the middle syllable of trisyllabic words, which is known to be a prosodically weak position (Chao, 1968; Shih, 1993), leading to much heavier contextual distortion than in the data used in the present study. Furthermore, the tonal information in Xu (1994) was further degraded by the voiceless initial consonants in the target syllable, which both hide and perturb the F_0 contours that were already quite short in duration due to the prosodically weak position (Xu, Xu, & Sun,

2003). What the findings of the present study demonstrate is that as long as the contextual distortion is not too severe and sufficient amount of F_0 movements toward the tonal target is available in the input (i.e., not hidden by voiceless consonants), a learning system can successfully derive the tonal categories without directly processing contextual information. It would be interesting in future investigations, of course, to explore how direct processing of contextual information can be used for helping the recognition of speech sounds that have been more severely distorted by contexts.

Finally, the present findings also have implications for another long-standing debate over the nature of speech perception, i.e., whether it is the auditory patterns (e.g., Diehl & Kluender, 1989) or articulatory gestures (Liberman & Mattingly, 1985) that are the distal objects of speech perception. While the auditory accounts may have difficulty explaining how variability with apparent articulatory sources can be effectively processed without referring to the articulatory movements, the motor theory may have difficulty explaining how infants who cannot yet speak can develop perceptual phonological categories that are articulatory in nature. The learning simulations in the present study suggest that by tracking the velocity profile of an acoustic parameter that closely reflects the underlying articulatory movements, variability both due to individual difference and contextual variations can be drastically reduced. And, the remaining variability, being articulatorily lawful, can be effectively handled by a neural network through unsupervised learning. This finding is reminiscent of the direct realist view of speech perception (Fowler, 1986) which postulates that the objects of speech perception are articulatory gestures as opposed to auditory properties in the form of distinctive features. The direct realist view also postulates that speech perception is done by tracking articulatory movements. As we have seen, the learned prototypical velocity profiles in the present study directly reflect movements toward underlying targets that are either static or dynamic in terms of both acoustic patterns and articulatory states. It is therefore conceivable that a further learning step for the infants is to derive those targets from categorized velocity profiles. Once stored in the brain, infants may then use those targets as articulatory goals when they babble and learn to speak themselves. This understanding therefore allows the possibility that speech production and perception are closely linked to each other but not necessarily always in lockstep.

5. Concluding remarks

Given that the speech input to infants is highly variable, and that infants are not typically told what the meaningful sounds are in a language, one of the greatest puzzles about human speech is how an infant can discover the sound categories of the ambient language from adult input. In the present study, we investigated the possibility that infants can derive phonetic categories directly from the time-varying acoustic signals produced by adults without having to extract abstract features from the signal. To this end, we explored the hypothesis of the Target Approximation model of tone production (Xu & Wang, 2001) that the consistency of lexical tones

produced in connected speech in a language like Mandarin lies in the continuous articulatory movement toward the underlying targets of the tones, as reflected in the F_0 trajectories during the syllable. We also explored the possibility that the velocity profiles ($D1$) represent more directly (than F_0) articulatory movements toward the underlying targets of the lexical tones, and as such they can significantly reduce the amount of variability due to speaker difference and tonal context. We tested these possibilities with a self-organizing topographical neural network using syllable-sized F_0 and $D1$ profiles as input. Although the debate persists in the field of language acquisition about the role of feedback during learning, our simulations demonstrate that guided feedback is not needed for the learning system to successfully derive the tonal categories. Testing results showed that while F_0 gave reasonably good performance, the prototypical $D1$ profile clusters developed through training yielded virtually perfect tone recognition without the help of any contextual information or pre-abstracted features. Further simulation showed that the learned $D1$ clusters, through additional learning, can be developed into even more ideal prototypes that have one-to-tone correspondence to the tones. These findings not only point to a possible way via which infants can develop phonetic categories through unsupervised learning, but also may lead to answers to various theoretical questions about language acquisition, speech perception and speech production.

Acknowledgements

Part of the results of the study was reported at the 149th meeting of the Acoustical Society of America, 2005 and the ISCA Workshop on Plasticity in Speech Perception, 2005. We thank the support of a FCAR(FQRSC) scholarship to the first author, the funding from SSHRC, NSERC and FQRSC to the second author, and the support from NIH Grant DC006243 to the third author.

Appendix A. The SOM algorithm

A.1. Architecture

The SOM maps a high-dimensional input space onto a discrete lower dimensional array of topologically ordered processing units. A 1-dimensional SOM is illustrated in Fig. 8 (adapted from Ritter & Schulten (1986)). The input and output layers are fully interconnected to each other. Output space N is a lattice on which units are labeled by a position vector r indicating their physical position on that lattice (filled dots on the vertical line). Input space X is mapped on output space N by a set of adaptive receptive field centers, or connection weights $w_r \in X$ (empty dots on horizontal lines) for which corresponds a typical $x_i \in X$. The subset of X closer to a unit's receptive field center than to any other w_r constitutes the receptive field of that unit (vertical bold lines). In the present study, a two-dimensional map of 10×10 units is used.

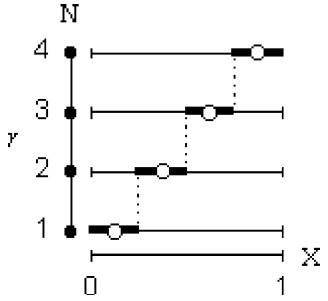

Fig. 8. Architecture of a one-dimensional SOM: linear array N of 4 output units r (filled dots), their receptive field center (empty dots) and receptive field (bold horizontal lines) for input space $X = [0, 1]$ (adapted from Ritter and Schulten (1986)).

A.2. Transmission rule

The transfer function of the network contains two steps. First, the distance between the receptive field center of each unit to that of input vector x_i is evaluated according to:

$$u_r = \left(\sum (x_i - w_r)^2\right)^{1/2}, \tag{7}$$

where u_r represents the net value of each unit r. The unit with the shortest Euclidean distance between w_r to reference input pattern x_i is selected to be the winner according to:

$$v = \min(u_r), \tag{8}$$

where v corresponds to the position of the winner, or Best-Matching-Unit (BMU). The net value is further transformed to yield the final response, given by the non-linear Gaussian function

$$\eta_r = \mathrm{Exp}[-([r - v]^2/\sigma)], \tag{9}$$

where η_r corresponds to each unit's activation. The Gaussian is peaked at v so the winning unit is the most activated ($\eta_v = 1$). Units falling into neighborhood radius σ get activated by means of afferent lateral activity, although to a lesser degree that depends on their position relative to the winner. The transmission rule can be conceived as a basic perceptual discriminative function that computes the distance between a perceived signal and a signal stored in a list of prototypes.

A.3. Learning rule

The SOM implements a regression algorithm for mapping an input distribution $P(x)$, $x_i \in X$, onto the output space. The lateral connections between output space nodes allow for topological ordering to be preserved in the map during the learning period. Receptive field centers w_r are adapted during a stochastic learning procedure

in which a random sequence of data points x_i is presented repeatedly for a predefined number of times. Each time an input vector is presented, the winning unit and its neighbors shift their receptive field center toward the data point according to

$$\Delta w_r = \alpha \cdot \eta_r (x_i - w_r), \tag{10}$$

where η_r is the value outputed by the transmission rule and α is the learning step size. The weight matrix is then updated according to

$$w_r(t+1) = w_r(t) + \Delta w_r. \tag{11}$$

The learning rule can be conceived as a basic perceptual learning function that transforms the internal organization to reflect the environment characteristics.

A.4. Initialization of the map

The weight matrix is initialized as follows. Each receptive field center is set to correspond to a linear trajectory of the form $ax + b$, where the slope $a = 0.00009$ and the intercept b ranges from 0.03 to 1. This yields a map in which the minimum and maximum values correspond to 0.03 and 1.99, which covers the input space (ranging from 0.54 to 1.95) without supposing a predefined number of categories. Other types of initialization schemes have been tried and made no difference in the final results, since whether the weights are bigger or smaller, they will respectively shrink or expand with the learning process to fit the input distribution.

References

Abramson, A. S. (1962). *The vowels and tones of standard Thai: Acoustical measurements and experiments* (Vol. 20). Bloomington, IN: Indiana University Research Center in Anthropology, Folklore, and Linguistics.

Abramson, A. S. (1978). Static and dynamic acoustic cues in distinctive tones. *Language and Speech, 21*(4), 319–325.

Andruski, J. E., & Ratliff, M. (2000). Use of phonation type in distinguishing tone: The case of Green Mong. *Journal of the International Phonetics Association, 30*, 39–62.

Bertoncini, J. R., & Mehler, J. (1981). Syllables as units in infant speech perception. *Infant Behavior and Development, 4*, 247–260.

Best, C. T. (1993). Emergence of language-specific constraints in perception of non-native speech: a window on early phonological development. In B. de Boysson-Bardies, S. Schonen, P. W. Jusczyk, P. McNeilage, & J. Morton (Eds.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 289–304). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Best, C. T. (1995). A direct realist perspective on cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research* (pp. 167–200). Timonium, MD: York Press.

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). American listeners' perception of nonnative consonant contrasts varying in perceptual assimilation to English phonology. *Journal of the Acoustical Society of America, 1097*, 775–794.

Bijeljac-Babic, R., Bertoncini, J., & Mehler, J. (1993). How do four-day-old infants categorize multisyllabic utterances?. *Developmental Psychology 29*, 711–721.

Bradlow, A. R., Pisoni, D. B., Yamada, R. A., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/ IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America, 101*, 2299–2310.

Brent, M., & Siskind, J. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition, 81*(2), B33–B44.

Chao, Y. R. (1933). Tone and intonation in Chinese. *Bulletin of the Institute of History and Philology, 4*, 121–134.

Chao, Y. R. (1968). *A grammar of spoken Chinese*. Berkeley, CA: University of California Press.

Chen, Y., & Xu, Y. (2006). Production of weak elements in speech—evidence from f0 patterns of neutral tone in standard Chinese. *Phonetica, 63*, 47–75.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.

Crottaz-Herbette, S., & Ragot, R. (2000). Perception of complex sounds: N1 latency codes pitch and topography codes spectra. *Clinical Neurophysiology, 111*, 1759–1766.

DeFrancis, J. (1984). *The Chinese language: Fact and fantasy*. Honolulu: University of Hawaii Press.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America, 27*(4), 769–773.

Diehl, R. L., & Kluender, K. R. (1989). On the objects of speech perception. *Ecological Psychology, 1*(2), 121–144.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14*, 3–28.

Fujimura, O. (2000). The C/D model and prosodic control of articulatory behavior. *Phonetica, 57*(2–4), 128–138.

Gandour, J. (1983). Tone perception in far eastern languages. *Journal of Phonetics, 11*, 149–175.

Gandour, J. (2000). Frontiers of brain mapping of speech prosody. *Brain and Language, 71*, 75–77.

Gandour, J., Potisuk, S., & Dechongkit, S. (1994). Tonal coarticulation in Thai. *Journal of Phonetics, 22*, 477–492.

Han, M. S., & Kim, K.-O. (1974). Phonetic variation of Vietnamese tones in disyllabic utterances. *Journal of Phonetics, 2*, 223–232.

Harrison, P. (2000). Acquiring the phonology of lexical tone in infancy. *Lingua, 110*, 581–616.

Howie, J. (1976). *Acoustical studies of Mandarin vowels and tones*. New York: Cambridge University Press.

Jacobson, R., Fant, G., & Halle, M. (1952). *Preliminaries to speech analysis: The distinctive features and their correlates*. MIT Acoustic Laboratory Technical Report 13. Cambridge, MA: MIT Press.

Jusczyk, P. W., & Derrah, C. (1987). Representation of speech sounds by young infants. *Developmental Psychology, 23*, 648–654.

Jusczyk, P. W., Pisoni, D., & Mullenix, J. (1992). Some consequences of stimulus variability on speech processing by 2-month-old infants. *Cognition, 43*, 253–291.

Karlgren, B. (1962). *Sound and symbol in Chinese*. Hong Kong: Hong Kong University Press.

Kaski, S., Venna, J., & Kohonen, T. (1999). Coloring that reveals high-dimensional structures in data. In *Proceedings of the 6th international conference on neural information processing* (pp. 729–734). Perth, Australia.

Kelso, J. A. S. (1984). Phase transitions and critical behavior in human bimanual coordination. *American Journal of Physiology: Regulatory, Integrative and Comparative, 246*, R1000–R1004.

Kelso, J. A. S., Saltzman, E. L., & Tuller, B. (1986). The dynamical perspective on speech production: data and theory. *Journal of Phonetics, 14*, 29–59.

Klein, D., Zatorre, R. J., Milner, B., & Zhao, V. (2001). A cross-linguistic PET study of tone perception in Mandarin Chinese and English speakers. *NeuroImage, 13*, 646–653.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43*, 59–69.

Kohonen, T. (1989). *Self-organization and associative memory*. Berlin: Springer.

Kohonen, T. (1995). *Self-organizing maps*. Berlin: Springer.

Krakow, R. A. (1999). Physiological organization of syllables: a review. *Journal of Phonetics, 27*(1), 23–54.

Kuhl, P. K. (1976). Speech perception in early infancy: the acquisition of speech sound categories. In S. K. Hirsh, D. H. Eldridge, I. J. Hirsh, & S. R. Silverman (Eds.), *Hearing and Davis: Essays honoring Hallowell Davis*. St-Louis: Washington University Press.

Kuhl, P. K. (1979). Speech perception in early infancy: perceptual constancy for spectrally dissimilar vowel categories. *Journal of the Acoustical Society of America, 66*, 1668–1679.

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development, 6*, 263–285.

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., & Miller, J. D. (1982). Discrimination of auditory target dimensions in the presence or absence of variation in a second dimension by infants. *Perception & Psychophysics, 31*, 279–292.

Laniran, Y. O., & Clements, G. N. (2003). Downstep and high raising: interacting factors in Yoruba tone production. *Journal of Phonetics, 31*, 203–250.

Liang, Z. A. (1963). Auditory perceptual cues in Mandarin tones. *Acta Physiologica Sinica, 26*, 85–91.

Liberman, A. M. (1970). Some characteristics of perception in the speech mode. *Perception and its Disorders, 48*, 238–254.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1–36.

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences, 21*, 499–511.

Maddieson, I., & Hess, S. (1986). 'Tense' and 'lax' revisited: more on phonation type and pitch in minority languages in China. *UCLA Working Papers in Phonetics, 63*, 103–109.

Massaro, D. W., Cohen, M. M., & Tseng, C. (1985). The evaluation and integration of pitch height and pitch contour in lexical tone perception in Mandarin Chinese. *Journal of Chinese Linguistics, 13*, 267–290.

Mattock, K. J. (2004). Perceptual reorganisation for tone: linguistic tone and non-linguistic pitch perception by English language and Chinese language infants. Unpublished doctoral dissertation, University of Western Sydney, Australia.

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

Menon, K. M., Rao, P. V., & Thosar, R. B. (1974). Formant transitions and stop consonant perception in syllables. *Language and Speech, 17*(1), 27–46.

Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *Journal of the Acoustical Society of America, 102*(3), 1864–1877.

Nelson, W. L. (1983). Physical principles for economies of skilled movements. *Biological Cybernetics, 46*(2), 135–147.

Ohde, R. N. (1988). Revisiting stop-consonant perception for two-formant stimuli. *Journal of the Acoustical Society of America, 84*(4), 1551–1555.

Pantev, C., Bertrand, O., Eulitz, C., Verkindt, C., Hampson, S., Schuierer, G., et al. (1994). Specific tonotopic organizations of different areas of the human auditory cortex revealed by simultaneous magnetic and electric recordings. *Electroencephalography and Clinical Neurophysiology, 94*, 26–40.

Peng, S.-h. (2000). Lexical versus 'phonological' representations of Mandarin Sandhi tones. In M. B. Broe & J. B. Pierrehumbert (Eds.), *Papers in laboratory phonology V: Acquisition and the lexicon* (pp. 152–167). Cambridge: Cambridge University Press.

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.

Pierrehumbert, J. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology VII* (pp. 101–140). Berlin: Mouton de Gruyter.

Prince, A., & Smolensky, P. (1993). In *Optimality theory: constraint interaction in generative grammar*. Rutgers University Center for Cognitive Science Technical Report 2.

Prom-on, S., Xu, Y., Thipakorn, B. (2006). Quantitative target approximation model: simulating underlying mechanisms of tones and intonations. In *The 31st international conference on acoustics, speech, and signal processing*, Toulouse, France.

Ritter, H., & Schulten, K. (1986). On the stationary state of Kohonen's self-organizing sensory mapping. *Biological Cybernetics, 54*, 99–106.

Seldon, H. L. (1985). The anatomy of speech perception: human auditory cortex. In A. Peters & E. G. Jones (Eds.). *Cerebral cortex* (Vol. 4, pp. 273–327). New York: Plenum Press.

Shen, X. S. (1990). Tonal coarticulation in Mandarin. *Journal of Phonetics, 18*, 281–295.

Shen, X. S., & Lin, M. (1991). A perceptual study of Mandarin tones 2 and 3. *Language and Speech, 34*(2), 145–156.

Shi, R., Morgan, J., & Allopenna, P. (1998). Phonological and acoustic bases for earliest grammatical category assignment: a cross-linguistic perspective. *Journal of Child Language, 25*(1), 169–201.

Shih, C. (1993). Relative prominence of tonal targets. In *Proceedings of the 5th North American conference on Chinese linguistics, Newark, Delaware*. University of Delaware: 36.

Sun, X. (2002). The determination, analysis, and synthesis of fundamental frequency (Doctoral dissertation, Northwestern University, 2002). *Dissertation Abstracts International B 63* (11), 5195.

Sundberg, J. (1979). Maximum speed of pitch changes in singers and untrained subjects. *Journal of Phonetics, 7*, 71–79.

Varfis, A. (1993). On the use of two traditional statistical techniques to improve the readability of Kohonen Maps. In *Proceedings NATO ASI workshop statistics neural networks*.

Wang, W. S. Y. (1967). Phonological features of tone. *International Journal of American Linguistics, 33*, 93–105.

Wang, W. S.-Y., & Li, K.-P. (1967). Tone 3 in Pekinese. *Journal of Speech and Hearing Research, 10*, 629–636.

Weijer, J. van de (1998). Language input for word discovery. Unpublished doctoral dissertation, University of Nijmegen, Nijmegen, The Netherlands.

Werker, J. F., & Lalonde, C. E. (1988). Cross-language speech perception: Initial capabilities and developmental change. *Developmental Psychology, 24*(2), 672–683.

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

Wessinger, C. M., Buonocore, M. H., Kussmaul, C. L., & Mangun, G. R. (1997). Tonotopy in human auditory cortex examined with functional magnetic resonance imaging. *Human Brain Mapping, 5*, 18–25.

Whalen, D. H., & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica, 49*, 25–47.

Xu, Y. (1994). Production and perception of coarticulated tones. *Journal of the Acoustical Society of America, 95*, 2240–2253.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics, 25*, 61–83.

Xu, Y. (1999). Effects of tones and focus in the formation and alignment of F0 contours. *Journal of Phonetics, 27*, 55–105.

Xu, Y. (2001). Sources of tonal variations in connected speech. *Journal of Chinese Linguistics, Monograph Series, 17*, 1–31.

Xu, Y. (2005). Speech melody as articulatory implemented communicative functions. *Speech Communication, 46*, 220–251.

Xu, Y., & Liu, F. (in press). Tonal alignment, syllable structure and coarticulation: toward an integrated model. *Italian Journal of Linguistics*.

Xu, Y., & Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech. *Journal of the Acoustical Society of America, 111*, 1399–1413.

Xu, Y., & Wang, Q. E. (2001). Pitch targets and their realization: evidence from Mandarin Chinese. *Speech Communication, 33*, 319–337.

Xu, Y., & Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics, 33*, 159–197.

Xu, C. X., Xu, Y., & Sun, X. (2003). Effects of consonant aspiration on Mandarin tones. *Journal of the International Phonetic Association, 33*, 165–181.

Yip, M. (2002). *Tone*. Cambridge: Cambridge University Press.

Zemlin, W. R. (1988). *Speech and hearing sciences: Anatomy and physiology*. Englewood Cliffs, NJ: Prentice-Hall.