Simulating the acquisition of lexical tones from continuous dynamic input

Bruno Gauthier and Rushen Shi
Département de psychologie, Université du Québec à Montréal, C.P. 8888, Succursale Centre-Ville, Montréal, Québec H3C 3P8, Canada
gauthier.bruno@courrier.uqam.ca, shi.rushen@uqam.ca

Yi Xu
Department of Phonetics and Linguistics, University College London, Room G32, Wolfson House, 4 Stephenson Way, London, NW1 2HE, United Kingdom
yi@phon.ucl.ac.uk

Abstract: Infants develop phonetic categories by simply being exposed to adult speech. It remains unclear, however, how they handle the extensive variability inherent to speech, and how they process multiple linguistic functions that share the same acoustic parameters. Across four neural network simulations of lexical tone acquisition, self-organizing maps were trained with continuous speech input of increasing variability. Robust tonal categorization was achieved by tracking the velocity profiles of fundamental frequency contours. This result suggests that continuous speech signal carries sufficient categorical information that can be directly processed, and that dynamic acoustic information can be used for resolving the variability problem.

PACS numbers: 43.71.An, 43.71.Ft, 43.70.Mn [JMH]
Date Received: December 6, 2006 **Date Accepted:** February 21, 2007

1. Introduction

Before reaching 1 year of age, infants have developed the ability to process speech sounds specific to their native language. Infants are born with general auditory mechanisms to process all speech sounds of human languages,[1,2] and later show a decline in sensitivity to non-native categories while narrowing in on the native language phonetic categories.[3–5] This development seems to be related to their sensitivity to the statistical distribution of speech sounds, as it was recently demonstrated that 6–8-month-old infants develop discrimination sensitivity corresponding to bi-modal distribution of VOT values in the stimuli after brief training.[6] However, the speech input to infants is much more variable than bi-modal distributions. Speech signal is continuous and dynamic, and steady-state patterns are rare.[7] There is extensive contextual variability due to coarticulation[8] and large amount of cross-speaker variations. Moreover, there are often more than two phonetic categories along any particular acoustic/articulatory dimension, and infants do not know in advance how many of them there are in a given language. It thus remains unclear how the early perceptual system can develop phonetic categories from such highly variable acoustic input.

Here we study with unsupervised neural networks whether phonetic categories can be discovered by directly processing continuous speech signals that contain different types of variability and competing linguistic functions. We simulate the acquisition of lexical tones in Mandarin. Mandarin has four tones for distinguishing words that can be identical in segmental compositions: High (H), Rise (R), Low (L), and Fall (F), which are carried by the fundamental frequency (F_0) of the vocal fold vibrations. Because tones typically involve a single primary acoustic dimension, namely, F_0, they are ideal for testing hypotheses about detailed mechanisms of phonetic acquisition.

The F_0 values of a tone, however, vary extensively due to at least three sources (Fig. 1). First, cross-speaker variability arises from differences such as age, gender, and idiosyncrasies (e.g., Refs. 9 and 10). Second, contextual variability arises from neighboring sounds affecting

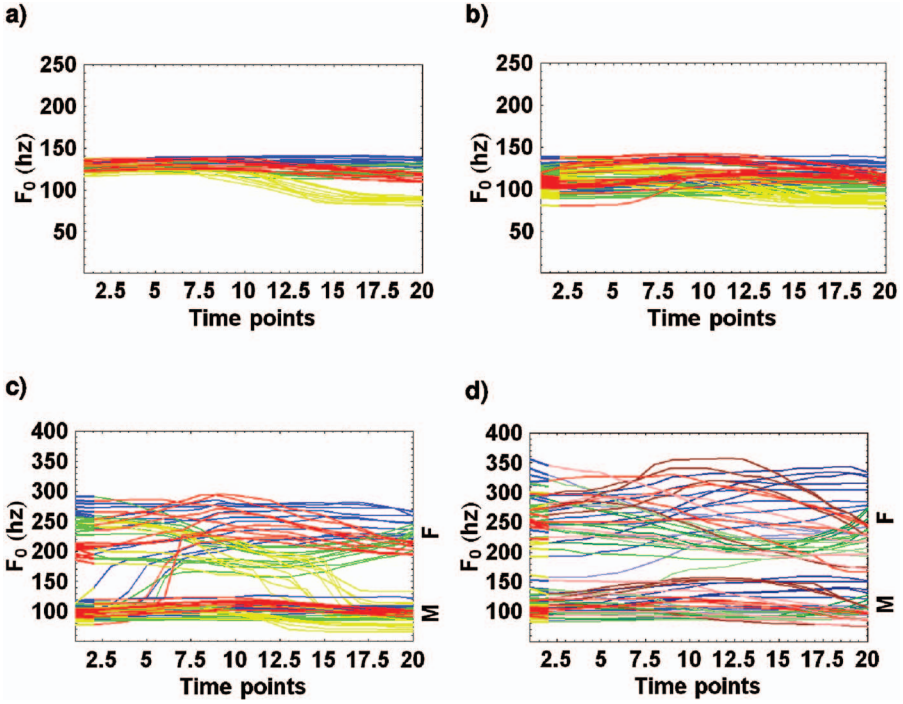

Fig. 1. Variability in tones. (a) F_0 (in hertz) of 40 repetitions of the four Mandarin tones (High, Rise, Low, Fall) by one male speaker with identical preceding tone (High) and identical focal status (neutral). (b) Contextual variability: F_0 of 80 repetitions of the four tones by one male speaker with different preceding tones (specified by syllable onset color) and identical focal status (neutral). (c) Contextual and speaker variability: F_0 of 80 repetitions of the four tones by one female and one male speaker with different preceding tones and identical focal status. (d) Contextual, speaker, and focal variability: F_0 of 80 repetitions of the four tones by one female and one male speaker with different preceding tones and variable focus (dark=on-focus, medium=neutral focus, pale=post-focus).

one another.[11] The F_0 pattern of any tone varies extensively due to the mechanical carryover influence of the preceding tone.[12] For example, the F_0 contour of a High tone following a Low tone somewhat resembles that of a Rise tone. Speech addressed to infants consists primarily of multiword utterances,[13] leading to considerable contextual variability. Finally, variability in F_0 comes from its use to not only distinguish words, but also to encode information such as focus (for emphasizing part of a message), which can introduce F_0 variations with magnitudes similar to those of tones.[14]

Despite the extensive variability, tonal perception appears early in infancy.[15] A key to infants' strategies may lie in the understanding of the mechanism of contextual variability; that is, despite the apparent variability, the talkers' articulatory strategy remains the same: to approach a constant tonal target starting from the syllable-initial F_0 left by the tone of the preceding syllable.[16] Such a strategy would result in velocity profiles that directly reflect the nature of the tonal targets. Velocity profiles have been shown to reveal the dynamics of skilled actions such as jaw movements during speech.[17] Moreover, taking the derivative of a curve leads to the removal of all its constant term(s), eliminating any overall height differences such as those due to cross-speaker variability. Thus, our hypotheses are (a) tonal categories can be discovered by processing syllable-sized continuous pitch movement patterns and (b) the velocity of F_0 (i.e., the first derivatives of F_0 patterns, henceforth D_1), better reveals the invariant properties of tones than F_0.

2. Method

In four simulations, self-organizing-maps [SOMs (Ref. [18])] were presented with learning material of increasing degrees of variability to assess their impact on tonal categorization. Self-organizing networks have been useful for characterizing the mechanism underlying various language acquisition tasks (e.g., Refs. 13 and 19) and for solving statistical pattern recognition problems. Given infants' sensitivity to distributional properties of speech signals,[6] the SOM is ideal for modeling the perceptual learning of phonetic categories. The SOM combines a map of topologically ordered processing units with a high-dimensional input space. Each output unit is connected to the input space by an adaptive weight vector, the dimensionality of which corresponds to that of input vectors. When an input token is presented to the network, the winning unit (i.e., with the shortest Euclidean distance to the input vector) is activated, and its connection weights are moved toward the data point by the learning rule. After training, the SOM is expected to reveal the structure of the data.

In the present study, the map units were arranged in a square topology, and initial connection weight values were arbitrarily assigned to cover only a small portion of the input space. The input space was formed of a set of continuous F_0 or D_1 vectors. In simulation 1, the input corpus contained 1800 exemplars produced by three adult male native Mandarin speakers (data from Ref. 12). Each stimulus corresponded to the first or second syllable of the word "mama" produced with varying tones in the middle of a carrier sentence that had either high or low pretarget F_0 offset and posttarget F_0 onset. In simulations 2, 3, and 4, tonal exemplars were from 3840 declarative sentences produced by eight native Mandarin speakers (data from Ref. 14). Sentences were formed of a subject, verb, and object and contained five syllables, each word corresponding to one or two consonant-vowel (CV) syllable(s), where C was a sonorant (/m,n/), except when the Low tone occurred on the fourth syllable, where C was /d/. The subject and object words were disyllabic and the verb was monosyllabic. The sentences were produced in various focus conditions: (a) neutral focus, (b) focus on word 1, (c) focus on word 2, and (d) focus on word 3 [e.g., "maomi mo maomi" (kitty touches kitty)] in response to the following wh-questions: "What is Kitty doing?" "Who is stroking Kitty?," "What is Kitty doing to Kitty?," "What is Kitty stroking?"). Since the tone on the first and last syllables was kept constant to High, these syllables were removed from the input corpus. The second, third, and fourth syllables contained varying tones (H, R, L, F on the second syllable, H, R, F on the third syllable, and H, L on the fourth syllable). Simulations 2–4 thus involved stimuli produced in all possible tonal contexts and focus conditions. Simulation 2 involved 1440 exemplars produced by a single speaker, and simulation 3 used 5760 tones produced by four male speakers. Simulation 4 involved the highest amount of variability, with 11 520 tones produced by four male and four female speakers in all tonal and focus conditions.

Input tokens to the network were 20-point vectors composed of equal-distanced discrete values from syllable-sized F_0 curves. The continuous F_0 contour was extracted by taking the inverse of every vocal period (for the detailed F_0 extraction procedure, see Ref. 14). F_0 input vectors were first transformed from hertz scale to semitone scale according to

$$F_0st = 12 \log_2 (F_0hz) \tag{1}$$

The velocity profiles of F_0 were generated according to

$$D_{1i} = (F_0st_{i+1} - F_0st_i)/(T_{i+1} - T_i) \tag{2}$$

where T represents time, which yields the discrete first derivatives of F_0. Each simulation presented one network with F_0 and one network with D_1 to compare the performance of both parameters. D_1 was expected to yield better results than F_0 in all simulations, and focused syllables were expected to be categorized better than postfocused syllables.

To compare the results obtained in different simulations, the size of a map was proportional to the input corpus size by a factor of about $\frac{1}{10}$ [e.g., 144 (12×12) units for Simulation 2 (1440 tokens)]. During training, half of the input tokens were randomly presented to the networks as whole vectors. The learning step size decreased linearly from 0.7 to 0.01. The neigh-

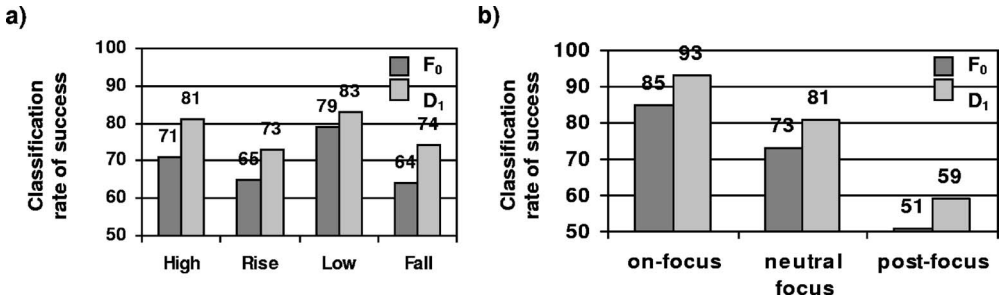

Fig. 2. Percent correct classification of F_0 (dark bars) and D_1 (pale bars) networks during simulation 4. (a) Results for tones High, Rise, Low, and Fall. (b) Results expressed as a function of focal status (on-focus, neutral focus, and post-focus).

borhood activation function included all units at the beginning, decreased exponentially, and only included the winning unit toward the end of training. The testing phase presented the training corpus and new input tokens to verify the networks' capacity to generalize to novel data. During testing, units that responded to a single category at least 68% of the time were labeled as that category. Units that responded to multiple tonal categories, none of which was dominant, were treated as noncategorical. Learning was assessed in terms of classification rate of success, i.e., the percentage of correctly classified input tokens.

To better understand the outcome of the categorization process, quantitative coloring of the trained maps was obtained by associating each tonal category with a distinct color produced with the CMYK color system (the High tone is represented by blue, i.e., a mix of cyan and magenta in the vector [1,1,0,0]; Rise=green [1,0,1,0]; Low=yellow [0,0,1,0]; Fall=red [0,1,1,0]). Each map unit was then associated to a four-dimensional vector, the values of which were specified according to the unit firing probabilities for each tonal class during testing [the fourth element K (black) was kept null]. As a consequence, units responding to a single tone are represented by a saturated color while noncategorical units are represented by "impure" colors. Learned categories are thus shown on the maps as regions of distinct colors.

3. Results

Overall, the networks trained with either F_0 or D_1 yielded above chance level performance, but those trained with D_1 performed better than those trained with F_0. In simulation 1 (input from different tonal contexts by three male speakers) a reasonably high rate of success was achieved for each tone with F_0 (H: 73%, R: 84%, L: 96%, F: 83%; mean: 84%, standard deviation: 9%), but D_1 yielded almost perfect categorization (H: 93%, R: 96%, L: 94%, F: 90%; mean: 93%, s.d.: 3%), replicating our previous study.[20] In simulation 2 (input from one speaker in different tonal contexts with variable focus, i.e., on-focus, neutral focus, and postfocus) the overall rate of success decreased relative to simulation 1 for F_0 (H: 68%,R: 69%, L: 88%, F: 63%; mean: 72%, s.d.: 11%) and for D_1 (H: 85%, R: 81%, L: 87%, F: 84%; mean: 84%, s.d.: 3%), suggesting that focus-induced variability is more detrimental to tonal categorization than cross-speaker (within-gender) variability. Simulation 3 tested the combined impact of the aforementioned sources of variability (input was from four male speakers in different tonal contexts with variable focus). Although the performance of both networks declined, D_1 (H: 82%, R: 76%, L: 77%, F: 79%; mean: 79%, s.d.: 3%) still performed better than F_0 (H: 64%, R: 59%, L: 75%, F: 70%; mean: 67%, s.d.: 7%). Finally, the results of simulation 4 (cross-gender, contextual, and focus induced variability) are shown in Fig. 2. Both F_0 (mean: 70%, s.d.: 7%) and D_1 (mean: 78%, s.d.: 5%) declined in performance relative to previous simulations, with D_1 still showing superiority. Most of the errors in simulation 4 involved postfocused elements [Fig. 2(b)]. The rate of success of the D_1 network for on-focus syllables remained excellent (93%), consistent with the expectation that focused elements should be perceptually salient.

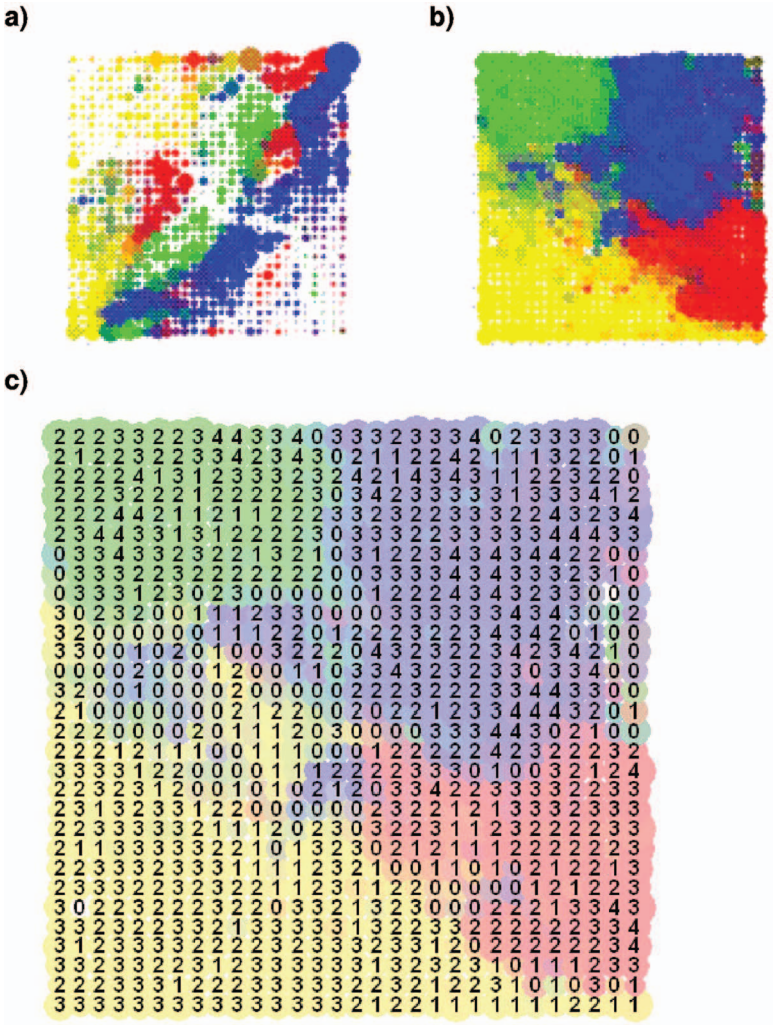

Fig. 3. Color maps of F_0 (a) and D_1 (b) networks in simulation 4 (High, Rise, Low, Fall). (c) Tonal context map for D_1 network, where categorical units are represented by the number of tonal contexts they are sensitive to (1 to 4; zeros correspond to noncategorical units).

Tonal color maps of simulation 4 are shown in Fig. 3. The F_0 color map [Fig. 3(a)] exhibits no clear tonal organization. Some tonal categories are distributed to multiple clusters. Activations of units during testing are uneven across the map: some units responded to many input tokens, while others to few or none (the larger a unit, the greater its firing probability). In contrast, the D_1 color map [Fig. 3(b)] shows four well-separated regions—each corresponding to one tone. Activations of units are even: each responding to a comparable number of test tokens. Verification of the response patterns of units shows that most categorical units responded to tones in multiple preceding contexts [Fig. 3(c)], confirming that each region indeed forms a single category and does not break down into subclusters based on tonal contexts (e.g., no subcluster for High tones preceded by other High tones). D_1 is thus much more powerful than F_0 in normalizing and categorizing the Mandarin tones.

4. Discussion and conclusion

The results of our simulations show that, despite extensive variability due to context, speaker, and competing linguistic functions (here tones and focus), a simple inductive learning mecha-

nism can extract tonal categories directly from continuous acoustic input without any supervision or feedback, assuming that F_0 contours have already been segmented into syllable-sized chunks by a separate mechanism.[21] We also find that velocity profiles as input for phonetic categorization are more robust than F_0 contours in handling the variability and revealing the invariant underlying phonetic targets. The effectiveness of using continuous dynamic patterns as input also eliminates the need for pre-extracting any summary properties as categorical cues, as such cues, if any, seem to operate implicitly rather than explicitly in the simulated learning process. Granted, no one has yet shown that infants, or even adults, have the ability to compute velocity profiles from acoustic signal. What the results of our simulations have demonstrated is that, if they did, the benefit would be enormous. It would thus be desirable for future research to look for direct evidence of neural processing of velocity.

Acknowledgments

Part of the results of the study were reported at the 151st meeting of the Acoustical Society of America, 2006. This work was supported by funding from SSHRC, NSERC, and FQRSC to the second author and supported in part by a NIH grant to the third author.

References and links

[1]R. N. Aslin, J. F. Werker, and J. L. Morgan, "Innate phonetic boundaries revisited," J. Acoust. Soc. Am. **112**(4), 1257–1260 (2002).
[2]P. D. Eimas, J. L. Miller, and P. W. Jusczyk, "On infant speech perception and the acquisition of language," in *Categorical Perception: The Groundwork of Cognition*, edited by S. Harnad (Cambridge University Press, New York, 1987), pp. 161–195.
[3]J. F. Werker and R. C. Tees, "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life," Infant Behav. Dev. **7**(1), 49–63 (1984).
[4]P. K. Kuhl, K. A. Williams, F. Lacerda, K. N. Stevens, and B. Lindblom, "Linguistic experience alters phonetic perception in infants by 6 months of age," Science **255**(5044), 606–608 (1992).
[5]P. K. Kuhl, E. Stevens, A. Hayashi, T. Deguchi, S. Kiritani, and P. Iverson, "Infants show a facilitation effect for native language phonetic perception between 6 and 12 months," Developmental Sci. **9**(2), F13–F21 (2006).
[6]J. Maye, J. F. Werker, and L. Gerken, "Infant sensitivity to distributional information can affect phonetic discrimination," Cognition **82**(3), B101–B111 (2002).
[7]J. S. Perkell and D. H. Klatt, *Invariance and Variability in Speech Processes* (Erlbaum, Hillsdale, NJ, 1986).
[8]W. J. Hardcastle and N. Hewlett, *Coarticulation: Theory, Data and Techniques* (Cambridge University Press, Cambridge, 1999).
[9]G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**(2), 175–184 (1952).
[10]J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**(5), 3099–3111 (1995).
[11]S. E. G. Ohman, "Coarticulation in VCV utterances: Spectrographic measurements," J. Acoust. Soc. Am. **39**(1), 151–168 (1966).
[12]Y. Xu, "Contextual tonal variations in Mandarin," J. Phonetics **25**, 61–83 (1997).
[13]R. Shi, J. L. Morgan, and P. Allopenna, "Phonological and acoustic bases for earliest grammatical category assignment: A cross-linguistic perspective," J. Child Lang **25**(1), 169–201 (1998).
[14]Y. Xu, "Effects of tone and focus on the formation and alignment of F_0 contours," J. Phonetics **27**, 55–105 (1999).
[15]K. J. Mattock, "Perceptual reorganisation for tone: Linguistic tone and non-linguistic pitch perception by English language and Chinese language infant," Unpublished doctoral dissertation, University of Western Sydney, 2004.
[16]Y. Xu and Q. E. Wang, "Pitch targets and their realization: Evidence from Mandarin Chinese," Speech Commun. **33**, 319–337 (2001).
[17]W. L. Nelson, "Physical principles for economies of skilled movements," Biol. Cybern. **46**(2), 135–147 (1983).
[18]T. Kohonen, *Self-Organizing-Maps* (Springer, Berlin, 1995).
[19]F. H. Guenther and M. N. Gjaja, "The perceptual magnet effect as an emergent property of neural map formation," J. Acoust. Soc. Am. **100**(2), 1111–1121 (1996).
[20]B. Gauthier, R. Shi, and Y. Xu, "Learning phonetic categories by tracking movements," Cognition **103**(1), 80–106 (2007).
[21]J. Bertoncini and J. Mehler, "Syllables as units in infant speech perception," Infant Behav. Dev. **4**(3), 247–260 (1981).