Proceedings of BUCLD 30: Proceedings of the 30th annual Boston University conference on language development, Boston, MA. Cascadilla Press: 202-212.

Variable Input and the Discovery of Lexical Tones in Infants: A Connectionist Approach

Bruno Gauthier[1], Rushen Shi[1], and Yi Xu[2]
[1]University of Quebec in Montreal, [2]University College London

1. Introduction – The tone system of Mandarin

One important task for infants in learning verbal communication is to discover the speech sounds of their ambient language. As revealed by the past 30 years of research in speech perception development, infants gradually become attuned to their native language phonetic categories during the second half of the first year of life. Most of the work in this area has focused on consonants (e.g., Werker & Tees, 1983, 1984; Werker & Lalonde, 1988; Best & McRoberts, 1989; Best, 1995) and vowels (e.g., Grieser & Kuhl, 1989; Kuhl, 1991; Polka & Werker, 1994). However, there is an emerging interest in the study of the acquisition of lexical tones, a type of contrast used by more than two thirds of the world's population (Yip, 2002).

Tone languages use distinct pitch patterns to distinguish word meanings. For example, the syllable /ma/ in Mandarin can mean 'mother' (High tone), 'hemp' (Rise), 'horse' (Low), or 'to scold' (Fall). The perception of pitch is highly correlated with the fundamental frequency (F_0) of the signal, which in turn reflects the vibration rate of the vocal folds. Although tone perception makes use of different phonetic markers such as duration and amplitude (Whalen & Xu, 1992) as well as phonation type (Maddieson & Hess, 1986; Andruski & Ratliff, 2000), F_0 is usually considered the primary acoustical cue for adult tone perception (Klein, Zatorre, Milner, & Zhao, 2001; Whalen & Xu, 1992). Figure 1 shows the F_0 patterns of the four Mandarin tones produced in citation form (data from Xu, 1997).

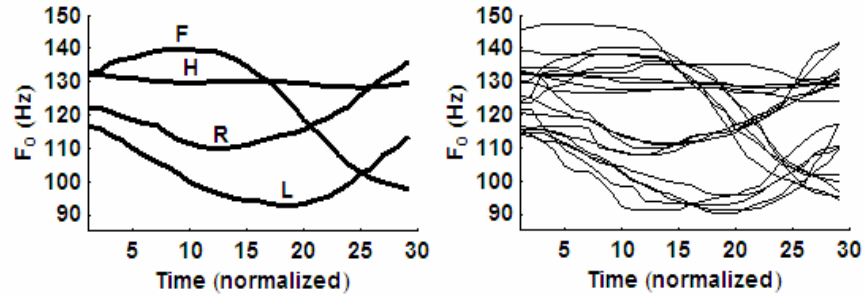

Figure 1. The lexical tones of Mandarin produced in citation form.

Each thick curve on the left panel represents the mean of five tokens produced by an adult male speaker. The thin curves on the right panel correspond to the individual tokens. As this graph shows, F_0 patterns seem to clearly distinguish the four tones. In everyday conversation, however, tones are generally not produced in isolation.

2. The problem of variability in tone production

There are many potential sources of variability in tonal realization, as discussed in detail by Xu (2001). In the present paper, we direct our attention towards two major sources: cross-speaker differences and contextual variations. The first one arises from the variability in length and thickness of speakers' vocal folds (Zemlin, 1988), resulting in pitch range variation. The second source is introduced by tonal contexts in connected speech (Shen, 1990; Xu, 1994; 1997). Much of the contextual variability in Mandarin is induced by the preceding tone; only a small portion of it comes from the following tone (Xu, 1997).

Input speech to infants conceivably also contains many sources of variability. For instance, their environment most likely includes multiple speakers. Furthermore, about 90% of infant-directed speech is produced in multi-word utterances (e.g., Weijer, 1998; Shi, Morgan &, Allopenna, 1998). Figure 2 shows a more realistic picture of the tonal input learners must be faced with in Mandarin.

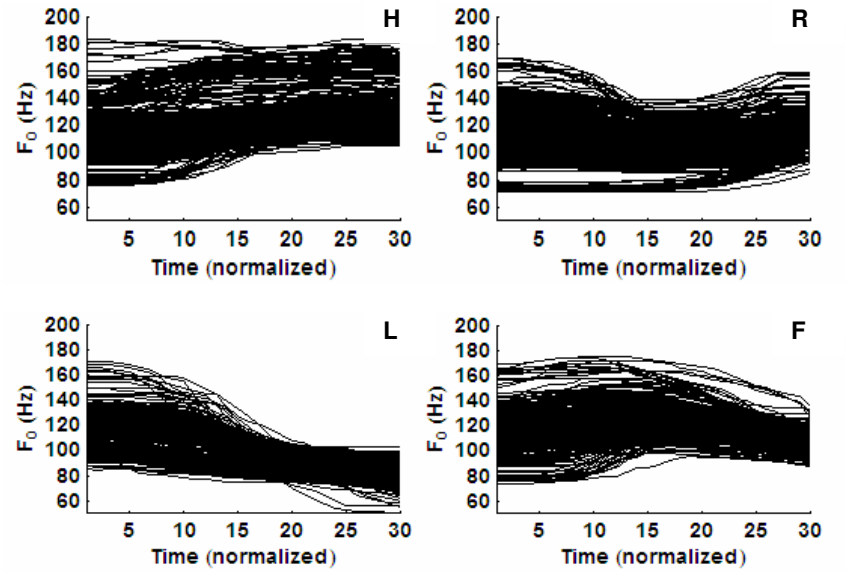

Figure 2. The Mandarin tones produced by 3 speakers in connected speech.

Each panel shows the tokens of a tonal class spoken by three speakers in connected speech (data from Xu, 1997). As can be seen, the perceptual system must deal with a substantial amount of variability. The two sources of variability described above result in extensive between-category overlap and within-category variability. This raises the question as to whether infants can learn the Mandarin tone system based solely on F_0 information.

3. Infant tone perception

Speech perception studies indicate that infants already demonstrate certain knowledge of tonal categories at the pre-verbal stage. In a recent study, Mattock (2004) showed that lexical tones undergo perceptual reorganization during the first year of life, as vowels and consonants do. Tonal perception is influenced as early as 9 months of age by the ambient tonemic system. Similarly, a study by Harrison (2000) indicates that infants learning a tone language start showing particular response patterns towards phonemic pitch variations as early as 6 months of age. Although no evidence yet exists regarding the exact age at which infants begin normalizing variability during tone perception, the fact that these studies used stimuli containing some amount of variability indicates that infants can grasp meaningful pitch variations quite early. In the present study, we explore *how* infants may achieve such normalization in order to learn their initial tonal categories.

4. Insights from tone production

Recent production work has shown that despite different sources of variability, F_0 contours of a tone all gradually converge over time to an asymptote that is characteristic of its underlying form: high-level for High, low-rising for Rise, low-level for Low and high-falling for Fall (Xu, 1997).

To account for these observations, Xu and Wang (2001) proposed the Target Approximation (TA) model of tone production. The model portrays the changing surface F_0 as resulting from different physical constraints imposed on the articulators during the implementation process. Tonal pitch variations are described as local asymptotic movements towards the underlying pitch targets, defined as simple linear functions. The targets can either be static or dynamic, and are respectively specified by relative pitch height (e.g., [high], [low]) and by both pitch velocity and relative height (e.g., [rise], [fall]).

The TA model makes an interesting prediction regarding the perception of tones. By assuming that tonal pitch variations always converge towards the underlying pitch targets, perception should have no difficulty in retrieving tonal identity from the speech signal, whether spoken in isolation or in connected speech. More specifically, the model predicts that it is possible to infer underlying pitch targets from the velocity profiles of F_0 movements. Velocity profiles (henceforth referred to as D_1) correspond to the first derivatives of F_0

contours and represent the instantaneous rates of change of the vocal folds vibration during tonal production.

The present study extends the TA model by proposing that D_1 may be important information that infants use to derive tonal categories early in language development.

5. Goal of the study

To test the hypothesis that a learner can use D_1 to derive the Mandarin tone system despite variability, we trained Self-Organizing-Maps (SOMs: Kohonen, 1982, 1995) with either continuous syllable-length F_0 patterns or their corresponding velocity profiles (i.e., D_1). In Simulation 1, the data for training and testing the networks contained both cross-speaker and contextual variability. In simulation 2, a second stage of learning was simulated for modeling further abstraction of the tonal categories. We expect that the efficiency of F_0 patterns for handling variability would be limited. On the other hand, D_1 information is expected to be a powerful cue for normalizing and categorizing the four Mandarin tones at both stages of learning (for details, see Gauthier, Shi, & Xu, submitted).

6. The SOM's algorithm

The SOM is a topographical neural network using unsupervised learning techniques for mapping a continuous high-dimensional input space onto a lower dimensional array of discrete processing units. The input distribution contains vector signals x_i elements of X. The neural map is a n-dimensional squared lattice of processing units, each labeled by a position vector r indicating its physical position on the lattice. The units are interconnected to their neighbors through lateral feedback and connected to the input space by a set of adaptive receptive field centers w_r, also called connection weight vectors, each of which corresponds to some typical x_i. The subset of X closer to one w_r than to any other is the receptive field of that unit.

The network processes an input vector in the following way. First, the distance between x_i to each w_r is computed according to: $u_r = (\text{Sum } (x_i - w_r)^2)^{1/2}$. The unit with the shortest Euclidean distance u_r is then selected to be the winner according to: $v = \min (u_r)$, where v corresponds to the position of the winning unit. Finally, the output of the network is given by the radial basis function: $n_r = \exp (- ((x_i - v)^2 / s))$, where n_r corresponds to the map's activation. The winning unit is maximally activated (to 1) while units' activation falling into radius s is a function of their distance from the winning unit.

During the learning process, each time an input vector is presented to the network, the winning unit and its neighbors shift their receptive field centers towards the data point according to: $D\ w_r = a\ .\ n_r\ (x_i - w_r))$, where a is the learning step size. The weight matrix is then updated according to: $w_r\ (t+1) = w_r\ (t) + D\ w_r$. The learning parameter a decreases linearly during training, from

0.05 to 0.001. The neighborhood radius *s*, which initially contains almost all units, decreases exponentially to eventually contain a single unit.

7. Simulation 1 – The learning of lexical tones

In this experiment, we test the efficiency of D_1 for normalizing and categorizing the four Mandarin lexical tones with naturalistic speech input, and then compare its performance with that of F_0.

The input corpus contains 1,800 exemplars of the four Mandarin tones produced in connected speech by three male speakers (data from Xu, 1997). Each tone corresponds to the first or second syllable of disyllabic 'mama' produced in carrier sentences varying in F_0 pre-target offset and post-target onset. Each stimulus is a 30-dimensional vector composed of equal-distanced discrete values taken from syllable-size time-normalized F_0 patterns (for the exact F_0 extraction procedure, see Xu, 1997). The F_0 profiles are first transformed from the Hertz to the Bark scale according to:

$$F_{0\,bk} = 7 \,.\, \mathrm{Log}\,(F_{0\,hz} \,/\, 650 + ((1 + (F_{0\,hz} \,/\, 650)^{\,2})^{\,1/2})$$

The velocity profiles are derived from F_0 by measuring the instantaneous velocity according to:

$$D_1 = 0.5\,(F_{0\,hz}\,(t+1) - F_{0\,hz}\,(t-1))$$

which yields input vectors of 28 dimensions representing the discrete first derivatives of F_0 patterns (Figure 3).

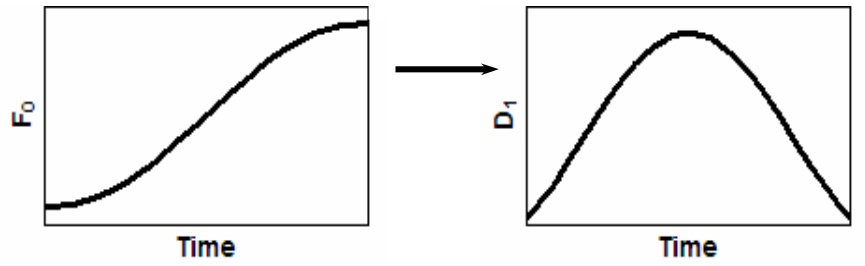

Figure 3. Illustration of F_0 to D_1 profile transformation of a High tone.

During the learning phase, half of the input corpus is used to train the network; 900 input vectors are presented randomly during 10 epochs, for a total of 9,000 presentations. The testing phase involves classifying the four tones correctly according to the output-coding scheme described in the next paragraph. A recall task thus presents the training corpus to the network, as well as a new set of 900 exemplars to verify its capacity to generalize to novel data. During recall, each token is assigned to the closest unit in terms of Euclidean distance (see the transmission rule in section 6).

The network is a squared array of 10 x 10 (100) processing units. Units that respond at least one time during the recall task are considered as operational units, while those that do *not* respond to any input vector are not considered any further in measuring the performance of the network. A unit is labeled as categorical if it is mostly sensitive to one input class. Units without a majority class are considered as confused units because they respond to multiple tones, none of which is dominant.

7.1. Results of Simulation 1

In this section, different measures are presented to compare the activity of F_0 versus D_1 networks during the testing phase. First, the overall performance measures on the trained maps indicate under which condition (F_0 vs. D_1) the data become most categorical. Then, more detailed measures on groups of map units reveal the distinctiveness and confusion patterns between the tonal classes.

The categorical error reflects the fraction of the map responding ambiguously during recall. It is expressed as the ratio of confused units to the total number of operable units. The results show that F_0 map contains 14% errors. In contrast, the D_1 map only contains 2% errors, indicating that most of the D_1 map units became category-specific after training while a larger fraction of the F_0 map contained confused units.

To qualitatively appreciate the outcome of the adaptation process, the phonotopic maps (Figure 4) assign to each unit the tone label corresponding to the majority class of that unit. The maps also indicate whether topologically ordered categories are present in the data and give information about the location of boundaries between each class. In the F_0 map (left), many confused units (multi-labeled units) can be observed, as well as widely spread confusion areas. In contrast, the D_1 map (right) shows a cleaner division of regions by classes, thus better representing the categories. These results suggest that the D_1 distribution contains more categorical information than does the F_0 distribution.

H	H	H	H	H	H	HR	R	H	H
H	H	H	H	H	H	H	H	HRF	F
HF	H	H	H	H	HRF	F	F	F	F
F	H	HRF	R	HRF	F	F	F	F	R
HF	H	HRF	R	HF	F	F	F	F	RF
HRF	HRF	R	R	F	F	F	F	F	F
R	R	R	R	R	R	R	F	F	L
R	R	R	R	RLF	R	R	L	L	L
R	RL	L	F	L	L	L	L	L	L
L	L	L	L	L	L	L	L	L	L

R	R	R	R	R	R	H	H	H	H
R	R	R	R	R	R	H	H	H	H
R	R	R	R	R	H	H	H	H	H
R	R	R	R	R	H	H	H	H	H
R	R	R	L	L	H	H	H	H	H
R	L	L	L	L	HF	F	F	F	F
L	L	L	L	F	F	F	F	F	F
L	L	L	L	F	F	F	F	F	F
RL	L	L	L	F	F	F	F	F	F
L	L	L	L	F	F	F	F	F	F

Figure 4. Phonotopic maps for F_0 (left) and D_1 (right) conditions.

Turning now to the detailed results, the rate of success for each tone corresponds to the ratio of correctly categorized tokens to the total number of test tokens of this category.

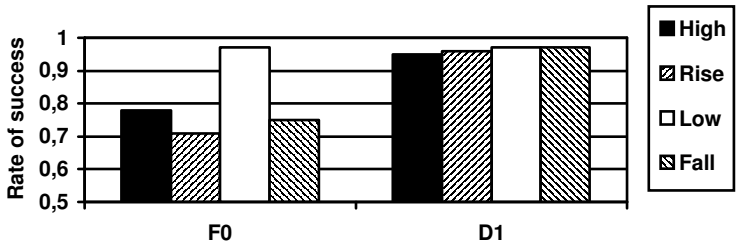

Figure 5. Rate of success for each tone in F_0 and D_1 maps.

Figure 5 shows that with F_0, only the Low tone is well recognized (97% of the time); the other tones show a mean success rate of 74% with 3.5% standard deviation. In contrast, the results from the D_1 condition indicate that every category shares a similar high rate of success (mean: 96%, sd: 1%). Together with the performance results and the phonotopic maps, the rate of success brings further evidence that D_1 better represents the four tonal categories than F_0 does.

One property of the neural maps trained with the SOM's algorithm is that they reflect the statistics of the input distributions. More specifically, the connection weight vectors tend to approximate the input space by extracting its important characteristics. To visualize the internal representation formed

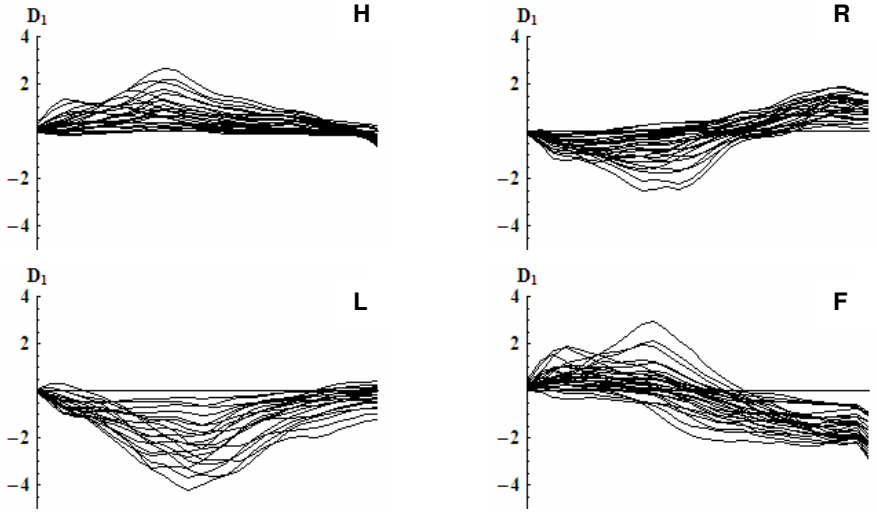

Figure 6. Velocity profiles formed by the D_1 map.

by a trained map, the connection weight vectors are usually projected onto the input space. Figure 6 displays the connection weight vectors of D_1 profiles for each group of map units associated to each tonal class.

The D_1 profiles show distinct movement patterns within each tonal category and demonstrate high consistency in terms of general direction of movement. The D_1 profiles of the static High and Low tones respectively speed towards high and low pitch values until they reach the middle of the syllable, and then gradually stabilize to their initial speed around the syllable offset. The D_1 profiles of the dynamic Rise and Fall tones show the same pattern until around the first third of the syllable, but then change course in the opposite direction, cross the zero speed line and continue to rise or fall until the end of the syllable. In this manner, the D_1 profiles seem to directly reflect the nature of the F_0 movements as characterized by the TA model (Xu and Wang, 2001). A learning system is evidently capable of using this information to normalize and categorize the four tones, as shown in the simulation.

8. Simulation 2 – A second learning stage

Since clusters corresponding to the four tones can be observed on the result map of Simulation 1, we can conclude that the simulated learning system formed distinct clusters after being trained with D_1 input. This seems to correspond to what infants may do in natural learning situations, i.e., forming tonal categories based on the clustering properties of the input data. Assuming that infants eventually develop a highly abstract system of tones, the next simulation examines whether a second learning stage could help the learner derive a more succinct representation of the four tonal categories.

The input corpus contains 72,000 tokens of F_0 or D_1 profiles corresponding to the responses generated in Simulation 1 every time an input token was fed into the 10 x 10 map during training. We now used a four-unit network (corresponding to the four tones) to test if the system could form exactly four categories correctly. During the learning phase, all 72,000 tokens are presented to the network in the same order they came out of the previous map. The testing phase is the same as in Simulation 1, using the whole input corpus.

8.1. Results of Simulation 2

The frequency matrices in Table 1 show how many tokens of each tonal class land on each unit during the recall task. As can be seen, each tone is more concentrated on a single unit on the D_1 map than on the F_0 map. For example, while the most often activated unit by the High tone in the F_0 map responds to 268 tokens (on a total of 480), 448 tokens of the same tone activate a single unit in the D_1 map. Another way to look at Table 1 is by doing a between-category comparison within each condition. For example, the most often activated unit by the Fall tone in the F_0 map (221) is activated an almost equal number of times by the High tone (268).

Table 1. Distribution of tonal tokens on map units.

	H	R	L	F
F_0	14 0 198 268	188 107 103 82	256 104 0 0	39 83 137 221
D_1	8 448 1 23	0 3 71 406	2 0 302 56	396 83 0 1

In contrast, the most often activated unit by the Fall tone (396) is activated by only a small number of High tones (8) in the D_1 map. These observations show that the D_1 map performs much better than the F_0 map in terms of classification of the four Mandarin tones, indicating that the learning system has successfully further abstracted the four tonal categories after the initial, less abstract learning phase.

Figure 7 shows the D_1 prototypes corresponding to the four Mandarin tones derived in Simulation 2. As can be seen, these seem to fit the descriptions of Figure 6 even better than the D_1 clusters shown there. This suggests that these more succinct D_1 profiles can actually represent most of the variants of the four Mandarin tones.

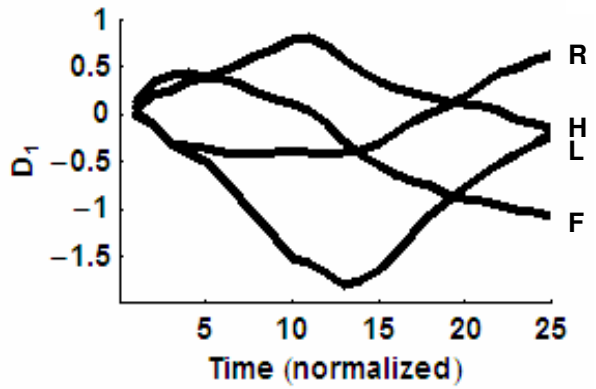

Figure 7. Prototypical velocity profiles formed during Simulation 2.

9. Discussion

The goal of this study was to test if the perceptual system of a naïve learner could derive the Mandarin tone system based on the velocity profiles of the fundamental frequency. To this end, we trained self-organizing networks using biologically plausible learning techniques with F_0 information versus D_1 profiles. The results show that D_1 profiles are indeed a superior source of information with which the learning system can categorize the four tones. After the initial

stage of successful category learning with D1, further mapping of the D_1 profiles resulted in abstraction of the four tonal categories onto four single units, whereas a similar attempt with F_0 failed.

Our findings suggest that naïve learners can successfully derive tonal categories from highly variable acoustical input by extracting underlying tonal targets based on perceived articulatory movements. Moreover, category formation may be achieved with raw acoustic patterns as input, without the need to first extract phonological features.

The claim that D_1 is the relevant information for deriving phonetic categories is consistent with recent advances in neurosciences. A study using positron emission tomography (PET) has located a functionally specialized area in the secondary auditory cortex involved in the processing of spectral changes such as the formant transitions of speech (Thivard, Belin, Zilbovicius, Poline, & Samson, 2000). This area may be responsive to changing acoustical information in general, including D_1. While the results of such studies, which were conducted with adult subjects, remain to be tested on infants, it seems plausible that infants process changes in F_0 contours at a very early age. Our study provides the first evidence that naïve learners can use velocity profiles to normalize and categorize tones in continuous speech input with high degree of variability. Naturally, it awaits future investigations to find out if a similar process actually happens during speech acquisition.

Acknowledgment

This study is supported by a FCAR scholarship to the first author, and grants from SSHRC and NSERC to the second author.

The first author would like to thank BUCLD for the Paula Menyuk Travel Award.

References

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.

Best, C. T. (1995). Learning to perceive the sound patterns of English. In C. Rovee-Collier and L.P. Lipsitt (Eds), *Advances in infancy research*. Norwood, NJ: Ablex.

Best, C. T., & McRoberts, G. W. (1989). Phonological influences on the perception of native and non-native speech contrasts. Paper presented at Biennial Meeting of the Society for Research in Child Development, Kansas City, MO, April 1989.

Gauthier, B., Shi, R., & Xu, Y. (Submitted). Learning phonetic categories by tracking movements. *Cognition*.

Grieser, D., & Kuhl, P. K. (1989). The categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology*, *25*, 577-588.

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

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. (1995). *Self-Organizing Maps*. Berlin: Springer.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics*, *50*, 93-107.

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.

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

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of non-native vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 421-435.

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

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.

Thivard, L., Belin, P., Zilbovicius, M., Poline, J.-B., & Samson, Y. (2000). A cortical region sensitive to auditory spectral motion. *Neuroreport*, *11* (13), 2969-2972.

Weijer, J. van de (1998). *Language input for word discovery*. Doctoral dissertation, University of Nijmegen, Nijmegen.

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

Werker, J. F., & Tees, R. C. (1983). Developmental changes across childhood in the perception of non-native speech sounds. *Canadian Journal of Psychology*, *37*, 278-286.

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.

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. (2001) Sources of tonal variations in connected speech. *Journal of Chinese Linguistics*, *monograph series #17*, 1-31.

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

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

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