Recognising tones by tracking movements – How infants may develop tonal categories from adult speech input

Bruno Gauthier[1], Rushen Shi[1], Yi Xu[2]

[1]Department of Psychology, University of Quebec in Montreal, Canada

[2] Department of Phonetics and Linguistics, University College London, United Kingdom

gauthier.bruno@courrier.uqam.ca, shi.rushen@uqam.ca, yi@phonetics.ucl.ac.uk

Abstract

Previous research has shown that the perception of speech in infants moves gradually from being language-general to being language-specific during the first year of life. Recently, it was found that infants learning a tone language begin to show particular response patterns to tones in their native language during their first year[1]. In this study we explore the relevance of tone production theory in understanding how the learning and the perception of Mandarin tones could have been accomplished despite highly variable speech input. The Target Approximation model [2] predicts that it is possible to infer underlying pitch targets from the manners of F_0 movements, for they may directly reflect the characteristics of intended goals. Using the production data of multiple speakers in connected speech from [3], we trained a self-organising neural network with both F_0 profiles and F_0 velocity profiles as input. The network's performance indicates that velocity profile distribution in the network formed distinct regions of clustering neighbourhoods representing each tone. The finding points to one way through which infants can successfully derive at phonetic categories from adult speech, namely, by extracting underlying phonetic targets based on information directly reflecting production.

1. Introduction

The tone system of Mandarin consists of four full tones, High (Tone 1), Rise (Tone 2), Low (Tone 3) and Fall (Tone 4) and a neutral tone [4]. The primary acoustic correlate of these tones is F_0 [5], also considered as the main cue in adult tone perception [6]. As other types of phonemes, Tones in tonal languages contain patterns of variability. For example, F_0 patterns produced by multiple speakers show a great amount of overlap for evident reasons. Moreover, tones produced in connected speech yield F_0 patterns with considerable between-category variability and large within-category overlap [7].

Many speech perception studies have focused on different aspects of the speech signal for finding invariants defining tonal categories [8,9,10,11]. Proposed solutions typically attempt to single out an acoustic parameter such as the height or slope of F_0 contours that remains constant for each tonal category. These proposals therefore all implicitly assume that some kind of preprocessing is done to derive these parameters before tonal categorization. Such preprocessing, however, seems even more difficult than the categorization itself, because it has to handle vast amount of variability in the speech signal [12]. The variability comes from two major sources: cross-speaker difference and context variation. For the adult listeners, it is imaginable that having preestablished categories may help the process of normalizing away the variability. For the prelinguistic infants, the task is much more difficult, because for them even the number of categories is unknown, not to mention the invariant parameter allegedly associated with each category. Nevertheless, recent findings suggest that some perceptual analysis of tonal variability must have already happened before the onset of tonal production [1,13,14]. Since speech input to infants consists primarily of multi-word utterances by multiple speakers [e.g., 15], tone learning must involve processes that can not only resolve the two types of variability, but also discover the number of tonal categories as well as their invariant characteristics.

Research on the nature of the variability in tone production has shown that much of the contextual variability in Mandarin is induced by the effect of the preceding tone, and in some contexts by the following tone [3]. It has also been shown that regardless of the preceding tone, the F_0 contours of the syllable associated with the tone all gradually converge over time to an asymptote that is characteristics of the underlying tone: high-level for High, low-rising for Rise, low-level for Low and high-falling for Fall [3].

To account for the contextual variability of tones, the Target Approximation (TA) model [2] characterises surface F_0 as asymptotic movements toward underlying pitch targets defined as simple linear functions. Targets can be either static or dynamic. Static targets are specified by relative pitch height (e.g., [high], [low]) and dynamic targets by a combination of relative pitch height and velocity of the pitch movement (e.g., [rise], [fall]). The TA model predicts that despite the variability, it is possible to infer underlying pitch targets from the velocity profile of F_0 movement (i.e., its first derivative, heterto referred to as D1), which represents the changes in fundamental frequency of the vocal folds in tonal production.

In the present study we test if tonal categories can be derived by the perceptual system of naïve learners despite the extensive speaker and contextual variability. We hypothesise that D1 of F_0 unifies the variable pitch trajectories of the same pitch target and distinguish them from those of other targets better than F_0 does. Given that infants begin voluntary vocal control of pitch changes from early stage of babbling, it is plausible that they use this information in discovering the intended underlying tonal targets.

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-organising topographical neural network. The Self-Organising-Feature-Map (SOM) [16] is a statistical pattern recognition device using unsupervised learning methods for discovering the structure of high dimensional data. The SOM maps a continuous input space onto a discrete lower dimensional array of topologically ordered processing units.

2.1. Description of the model

Architecture. The input space X and the output layer are fully interconnected by a set of adaptive receptive field centres $w_r \in$

X. The output is a lattice on which units are labelled by a position vector r indicating their physical position on that lattice. The subset of X closer to a unit's receptive field centre than to other w_r constitutes the receptive field of that unit.

Transmission rule. The distance between the receptive field centre of units to that of input vector is evaluated according to:

$$u_r = (\Sigma (x_i - w_r)^2)^{1/2} \quad (1)$$

where u_r represents the net value of 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:

$$v = \min (u_r) \quad (2)$$

where v corresponds to the position of the Best-Matching-Unit (BMU). The net value is further transformed to yield the final response, given by the nonlinear Gausian function:

$$\eta_r = \text{Exp}(- ((v - r)^2 / \sigma)) \quad (3)$$

where η_r corresponds to units' activation. Radius σ activates neighbouring units by means of afferent lateral activity.

Learning rule. A regression algorithm maps the input distribution P(x), $x_i \in$ X, onto the output space. Each time an input vector is presented, the BMU and its neighbours shift their receptive field centre towards the data point according to:

$$\Delta w_r = \alpha . \eta_r (x_i - w_r) \quad (4)$$

where α is the learning step size. The weight matrix is then updated according to:

$$w_r (t+1) = w_r (t) + \Delta w_r \quad (5)$$

2.2 Simulations

Input coding. The input corpus [3] contains 1800 exemplars of the four Mandarin tones produced in connected utterances by three adult male speakers. Each stimulus corresponds to the first or second syllable of disyllabic sequence 'mama' produced in the middle of four carrier sentences which differ in high and low pre-target offset and post-target onset. Each input token is a 30-data-point vector composed of equally distanced discrete values taken from a syllable-length time-normalised F_0 curve (for details, see [3]). The data are first transformed from Hertz to the Bark scale according to:

$$F_{0\,bk} = 7 . \text{Log}(F_{0\,hz}/650+((1+ (F_{0\,hz}/650)^2)^{1/2}) \quad (6)$$

For D1 simulation, the first derivatives of F_0 are :

$$D1 = 0.5 (F_{0\,hz} (t+1) - F_{0\,hz} (t-1)) \quad (7)$$

which yields input vectors of 28 dimensions.

Learning phase. The training corpus contains 900 stimuli (half of the input corpus) randomly presented to the network for 1000 times. Each time the neighbourhoods on the map are shifted to better fit the data. After learning, the map represented by the connection matrix is saved.

Testing phase. During the recall task, new exemplars are used to verify the network's capability to generalise to novel data. The trained network assigns each input pattern from a new set of 900 tokens (the other half of the input corpus) to a single unit with the transmission rule.

Output coding. The networks are squared arrays of 10 x 10 units, each being tuned to a particular subset of input patterns after training. Units which fire at least once during recall are considered operable units. If a unit never fires, it is non-operable. The number of activations of each unit for each tone category is indexed into four tone frequency matrices. Units with firing probability above 68% to a tone are considered as categorised and labelled with that particular tone. Units without such amajority class respond to multiple tones and are considered as ambiguous units.

2.3. Measures

The *categorical error* represents the proportion of the network responding to more than one class, i.e., the number of ambiguous units on the number of operable units. The *classification error* is the probability of the network to respond ambiguously during recall. Each test token landing on an ambiguous unit counts as an error, thus the classification error is the number of error tokens divided by the total number of input tokens in the testing corpus. For visualising the results, the *phonotopic map* [17,18] assigns each unit a label corresponding to the majority class of that unit

The between-category assessment of each condition is expressed in terms of confusion pattern between tones and is presented in the form of *confusion matrices*. Derived from these tables, the *rate of success* for each tone corresponds to the number of correctly categorised tokens divided by the total number of test tokens of this category.

3. Results

Global results. The global results indicate if topologically ordered categories are present in the data. Table 1 shows categorisation errors (column 2) and classification errors (column 3) for F_0 and D1. Both errors are larger for F_0 than for D1, indicating that after training, most of the D1 map units became category-specific while a larger portion of the F_0 map contained ambiguous units.

Table 1: Categorisation and classification errors.

	Global measures	
	Categorisation error	Classification error
F_0	0.20	0.22
D1	0.03	0.03

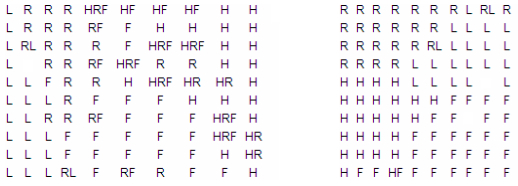

Figure 1: Phonotopic maps for F_0 (left) and D1 (right).

As shown in Figure 1, ambiguous units (multi-labelled units) form widely spread confusion areas on the F_0 map. In contrast,

the D1 map shows a cleaner division of regions by classes, thus better representing the categories. These results suggest that the D1 distribution contains more categorical information than does the F_0 distribution.

Detailed results. Table 2 presents a confusion matrix where the rows correspond to the speakers' intended targets and the columns to the majority class of the map units. The last column shows the number of tokens activating ambiguous units for each category. For example, in the F_0 condition, of the 240 intended High Tones, 162 were classified correctly as the High units, 6 misclassified as Rise units and 72 landed on ambiguous units.

Table 2 shows that 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 2 (Rise) was mostly misclassified as Tones 1, 3 and 4, and Tone 4 as Tones 1 and 2. The D1 condition shows only a single misclassification pattern between Tones 1 and 4.

Table 2: Between-category confusion pattern.

		Confusion matrix				
		H	R	L	F	Ambiguous
F_0	H	162	6	0	0	72
	R	**11**	157	**8**	**6**	58
	L	0	0	172	0	8
	F	**11**	**12**	1	160	56
D1	H	219	1	2	**8**	10
	R	2	226	3	0	9
	L	0	0	172	1	7
	F	**9**	0	1	224	6

With D1, the number of tokens assigned to the corresponding majority class is overall higher than in F_0, as shown by each matrix diagonal, although the Low Tone in F_0 behaves differently. The rate of success of the Low Tone corresponds to 172 correctly classified Low tokens divided by the total of Low tokens (180), thus 0.96. The mean success rate of F_0 for the three other Tones is 66% with one standard deviation.

In contrast, the results from the D1 condition indicate that every category shares a similar high rate of success (mean: 94%, sd: 2). Together with the confusion pattern results, the rate of success brings further evidence that D1 better represents the four tonal categories.

4. Discussion

The simulations compared the performances of F_0 and D1 as input to a topographical neural network consisting of 100 receptive units. The performance of D1 turned out to be far superior to that of F_0. D1 yields an almost perfect separation of the four tonal categories despite the multiple sources of variability in the speech input.

To understand why D1 is so much more effective than F_0, Figure 2 displays the prototypical F_0 and D1 profiles developed during training in the simulations 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 general directions: up and down. The D1 profiles, in contrast, display high consistency in terms of general direction of movement, and differ within each tonal category mostly in magnitude of the movement.

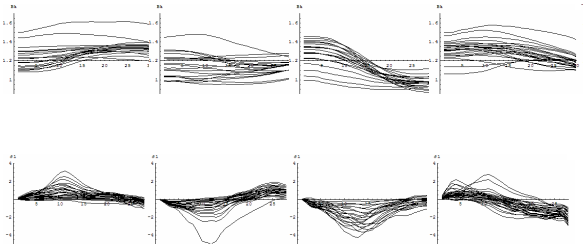

Figure 2: Prototypical profiles for F_0 (above) and D1 (below) for categorised units of the Tones 1, 2, 3 and 4.

Static tones. The consistent D1 profiles seem to directly reflect the nature of the F_0 movements as characterized by the Target Approximation model [2] and by the velocity profiles of movements proposed by Nelson [19]. In the static tones, most High profiles increase their speed from 0 towards a positive value, reach peak velocity around the 10-12th time point and finally slow down towards the initial speed of 0 near the end of the syllable. The Low profiles show almost mirror images of the High profiles, i.e. with speed increasing towards negative values from 0, reaching a valley around the 10-12th time point, and then decreasing towards 0 again near the end of the syllable. Such unimodal velocity profiles fit the definition of a single movement given by [19]. A single movement is one that starts from one location 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 towards a single static F_0 height. The positive velocity profiles during High correspond to movements towards an above-average pitch height, and the negative velocity profiles during Low correspond to movements towards a below-average pitch height.

Dynamic tones. 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 towards a negative/positive value, although for a shorter period of time. But instead of continuing with the initial direction, the D1 profiles reverse their direction, cross the zero speed line and continue to increase until near the end of the syllable. In other words, the Rise/Fall velocity profiles indicate rapid initial F_0 movement towards a relatively low/high F_0, followed by another movement in the opposite direction towards the zero line, thus indicating a movement toward an initial static height per [19]'s definition. But the movements afterwards no longer fit that definition. Rather, the fact that D1 reaches a high (positive or negative) value near the end of the syllable in Rise and Fall suggests that the high velocity itself is the final goal of these tones. In other words, the targets of these tones are dynamic, i.e., with a likely simple linear function as their goal, as is assumed in the Target Approximation model [2].

The present findings have implications for a long-lasting debate over the nature of speech perception, i.e., whether it is

the acoustic patterns or articulatory gestures that are the distal objects of speech perception. While the auditory accounts have difficulty explaining how variability with apparent articulatory sources can be effectively processed without referring to the articulatory movements, the motor theory accounts have difficulty explaining how infants who cannot yet speak can develop perceptual phonological categories that are articulatory in nature. The data used in the present study suggest that one of the first problems infants have to solve is how to handle the large magnitude of variability that comes from either an idiosyncratic source or an articulatory source. The learning simulations that we conducted suggest that by tracking the velocity of articulatory/acoustic movements, variability from both sources is drastically reduced; the remaining variability, being articulatorily lawful, can be effectively handled by a neural network through unsupervised learning. On the other hand, the velocity profiles being tracked can make sense only when they are viewed as stemming from movements toward underlying targets that can be defined as targets that are either static or dynamic in terms of both acoustic patterns and articulatory states. It is therefore imaginable that a further learning step for the infants is to derive those targets from categorized velocity profiles like those shown in Figure 2. Once stored in the brain, those targets may be used by infants as articulatory goals when they babble and learn to speak themselves. This understanding can therefore provide a possible explanation as to why perception well precedes production in language acquisition.

5. Conclusion

Given that the speech input to infants is highly variable, one of the greatest puzzles about human speech is how infants discover the sound categories of the ambient language. Based on the Target Approximation model of tone production [2], we hypothesised that the velocity profiles (D1) represent more directly than F_0 profiles articulatory movements towards the underlying pitch targets of the lexical tones, and as such it can significantly reduce the amount of variability due to speaker difference and tonal context. We further hypothesized that naïve learners such as infants can use D1 information to develop tonal categories through unsupervised learning. We tested these hypotheses with a self-organising topographical neural network using both F_0 and D1 profiles as input. Testing results showed that not only was D1 far superior to F_0 for developing tonal categories, but also the prototypical D1 profile clusters developed through training yielded virtually perfect tone recognition without the help of any contextual information. These findings not only point out a possible way via which infants can develop phonetic categories through unsupervised learning based on adult input containing large amount of variability, but are also pertinent to our understanding of the link between speech perception and production in general.

6. Acknowledgements

We thank the support of a FCAR scholarship to the first author, the funding from SSHRC and NSERC grants to the second author, and the support from NIH Grant DC03902 and NIH Grant DC006243 to the third author.

7. References

[1] Mattock, K. J., *Perceptual Reorganisation for Tone: Linguistic Tone and Non-linguistic Pitch Perception by English Language and Chinese Language Infants*, Ph.D. Dissertation, University of Western Sydney, 2004.

[2] Xu, Y. and Wang, Q. E., "Pitch targets and their realization: Evidence from Mandarin Chinese", *Speech Communication, Vol. 33, 2001, p 319-337.*

[3] Xu, Y. "Contextual tonal variations in Mandarin", *Journal of Phonetics, Vol. 25, 1997, p 61-83.*

[4] Chao, Y. R., *A Grammar of Spoken Chinese*, University of California Press, Berkeley, CA, 1968.

[5] Howie, J., *Acoustical Studies of Mandarin Vowels and Tones*, Cambridge University Press, New York, 1976.

[6] Klein, D., Zatorre, R.J., Milner, B. and Zhao, V., "A cross-linguistic PET study of tone perception in Mandarin Chinese and English speakers", *NeuroImage, Vol.13, 2001, p 646-653.*

[7] Shen, X.S., "Tonal coarticulation in Mandarin", *Journal of Phonetics, Vol. 18, 1990, p 281-295.*

[8] Abramson, A. S., "Static and dynamic acoustic cues in distinctive tones", *Language and Speech, Vol. 21, no 4, 1976, p 319-325.*

[9] Gandour, J., "Tone perception in far eastern languages" *Journal of Phonetics, Vol. 11, 1983, p 149-175.*

[10] Massaro, D.W., Cohen, M.M. and Tseng, C.-Y., "The evaluation and integration of pitch height and pitch contour in lexical tone perception in Mandarin Chinese", *Journal of Chinese Linguistics, Vol. 13, no 2, 1996, p 267-290.*

[11] Shen, X.S. and Lin, M., "A perceptual study of Mandarin tones 2 and 3", *Language and Speech, Vol. 34, no 2, 1991, p 145-156.*

[12] Perkell, J. S. and Klatt, D. H. (Eds.), *Invariance and Variability of Speech Processes*, LEA, Hillsdale, NJ, 1986.

[13] Harrison, P., "Acquiring the phonology of lexical tone in infancy", *Lingua, Vol. 110, 2000, p 581-616.*

[14] Hua, Z. and Dodd, B, "The phonological acquisition of Putonghua (Modern Standard Chinese)", *Journal of Child Language, Vol. 27, 2000, p 3-42.*

[15] Shi, R., Morgan, J. L. and Allopenna, P., "Phonological and acoustic bases for earliest grammatical category assignment: A cross-linguistic perspective", *Journal of child language, Vol. 25, 1998, p 169-201.*

[16] Kohonen, T., "Self-organized formation of topologically correct feature maps", *Biological Cybernetics, Vol. 43, 1982, p 59-69.*

[17] Kohonen, T., *Self-Organization and Associative Memory*, Springer, Berlin, 1989.

[18] Kohonen, T., *Self-Organizing Maps*, Springer, Berlin, 1995.

[19] Nelson, W. L., "Physical principles for economies of skilled movements", *Biological Cybernetics, Vol. 46, 1983, p 135-147.*