Mandarin tone production can be learned under perceptual guidance — A machine learning simulation

Hanyi Meng, Chien-Tzu Chen, Yue Chen, Zirui Liu, Yi Xu
University College London
hanyi.meng.20@ucl.ac.uk; chien-tzu.chen.21@alumni.ucl.ac.uk; yue.chen.1@ucl.ac.uk; zirui.liu.17@ucl.ac.uk; yi.xu@ucl.ac.uk

ABSTRACT

Intuitively, speech production can be learned by imitating proficient speakers in language acquisition. But a recent computational simulation has shown that learning to produce English words can be achieved under the guidance of speech perception, without direct mimicry. In this study, we tested whether similar perception-guided learning also applies to Mandarin tone acquisition. We used PENTAtrainer, a pitch modelling tool to simulate learners' tone articulation, and a trained tone recognizer to simulate tone perception. Three learning methods with different optimization objectives were tested: 1) closeness of fit of f_0 contours, 2) tone recognition by an automatic tone recognizer, and 3) tone recognition plus minimization of mean f_0 difference at the initial learning phase. The results show that method 3 achieved the best learning outcome as evaluated by the tone recognizer and human listeners. Perception-guided tone learning is therefore shown to be effective if learners' exploration range can be reduced first.

Keywords: speech perception, vocal tone learning, computational modelling, speech synthesis

1. INTRODUCTION

It is still unclear how children acquire language spontaneously without explicit adult instructions. A popular idea is that they do it through imitation [16, 23, 24]. However, it has been difficult to computationally simulate such imitative learning [11, 20, 21, 25]. And a major source of difficulty has been the speaker normalization [13] or correspondence [5] problem. For example, because children's vocal tracts are much shorter than adults' [10, 27], their formants are much higher and more dispersed, making it difficult for children to directly imitate adult speech. However, recently it has been demonstrated computationally that this correspondence problem can be largely solved by using speech perception as a guide in production learning [15, 26, 34]. The production of simple English words with high intelligibility can be learned this way without directly imitating any specific utterances.

The effectiveness of simulating perception-guided vocal learning raises the question of whether tone learning can be simulated in a similar way. Theoretically, this is conceivable, and in fact should be easier since tones mainly involve a single acoustic dimension, i.e., fundamental frequency (f_0), and tones have been successfully modelled with PENTAtrainer, a Praat-based prosody modelling tool [31]. However, perception-guided learning with only a single acoustic dimension could also present problems due to the lack of cross-reference to other parameters.

This study is a preliminary test of perception-guided tone learning. But a strategy slightly different from [15, 26, 24] is applied. First, the acoustic imitation was simulated by an algorithm that optimized the matching of f_0 contours across a whole corpus consisting of many utterances produced by multiple speakers (including both males and females). Second, data from the same corpus were used to assess the learning outcome of both imitative and perception-guided learning, thus minimizing confounding in comparison.

2. METHOD

2.1. Corpus

The Xu1999 corpus used in this project was originally recorded for an experimental study of tone and focus in Mandarin [28], which consisted of utterances produced by four female and four male speakers. The utterances were five-syllable sentences composed of three words (two disyllabic and one monosyllabic), as shown in Table 1. As can be seen, the second, third and fourth syllables have varying tones, while the tone of the first and last syllable is always H. The speech was fluent, with a speech rate of roughly five syllables/s.

Word 1			Word 2			Word 3		
HH	māomī	'kitty'	H	mō	'touches'	HH	māomī	'kitty'
HR	māomí	'cat-fan'	R	ná	'takes'	LH	mǎdāo	'sabre'
HL	māomǐ	'cat-rice'	F	mài	'sells'			
HF	māomì	'cat-honey'						

Table 1: Tone patterns and corresponding sentences used as recording material. H, R, L, and F represent high, rising, low, and falling tones, respectively [28].

The sentences in the full corpus also varied in focus conditions: initial, medial, final and no focus [28]. The present study only used the neutral focus sentences, however, to simulate tone learning only.

The corpus was divided into a training set consisting of 6 of the 8 speakers, 3 males and 3 females, and a testing set consisting of 2 speakers (1 male, 1 female).

2.2. Modelling tool

The computational tool was a special-purpose version of PENTAtrainer [31] — an interactive Praat [3] script for modelling speech prosody. PENTAtrainer models tone and intonation by combining built-in articulatory dynamics (target approximation) [22, 33], parallel encoding [29], and global stochastic learning (simulated annealing [14]) [31]. The original version has been shown to generate intelligible and natural-sounding tone and intonation by optimizing f_0 contour fitting [31]. The f_0 fitting can be viewed as a form of learning by imitation, as it tried to maximize the similarity between the learned and the original f_0 contours.

The special-purpose version of PENTAtrainer used in this study included two additional learning methods, *learning by optimizing tone recognition*, and *learning by optimizing tone recognition and minimizing mean f_0 difference* (deltaf_0). The computationally intensive learning task was run by a Python executable called by the Praat script. The tone recognizer, also called by the Praat script, was a support vector machine (SVM) trained by the scikit-learn package [6] in Python with syllable-sized f_0 contours as input data. The trained model was able to recognize both tone and focus with high accuracy [7, 8]. In learning method 3, deltaf_0 was the utterance-wide mean f_0 difference between each original and synthetic contour, and it added a fraction of weight to the tone recognition error in the coarse-tuning phase of the learning:

$$e = 0.9\, e_r + 0.1\, d \qquad (1)$$

where e_r = 10 (1 – recognition rate [0,1]), and $d = f_{0orig} - f_{0orig}$, where f_{0orig} and f_{0orig} were utterance-wide mean f_0 of original and synthetic tones, respectively.

The coarse-tuning phase, consisting of the first 450 of the total 750 training iterations, optimized all tonal targets at once in each iteration, while the fine-tuning phase optimized each parameter (height, slope and strength [31]) of each tonal target at a time.

2.3. Procedure

The experiment proceeded in 5 steps:

Step 1. Training the tone recognizer on all the neutral-focus utterances in the Xu1999 corpus. The overall post-training recognition rate was 94.7%.

Step 2. Running PENTAtrainer in the training set with three learning methods, each repeating five times. The main simulated annealing parameters used were: *iteration* = 750, *learning rate* = 0.1, *starting temperature* = 700, and *reduction factor* = 0.98.

Step 3. Averaging the pitch targets of each tone learned from the five runs of each learning method to obtain three sets of tone targets.

Step 4. Running PENTAtrainer in the testing set with the mean tone targets to a) generate f_0 contours that were input to the automatic tone recognizer for tone recognition, and b) resynthesize all the utterances in the testing set with the model-generated f_0 contours.

Step 5. Playing the resynthesized utterances to listeners for perceptual tone identification and judgment of naturalness.

For step 5, the stimuli contained 320 recordings from the testing set, which were divided into four conditions: a) original recordings, b) recordings resynthesized with parameters learned from f_0 fitting, c) recordings resynthesized with parameters learned from recognition only, and d) recordings resynthesized with parameters learned from recognition+deltaf_0. The tones to be identified belonged to the second syllable, which was always /mi/, c.f. Table 1. Syllables in other positions carried fewer tones, varied in segmental compositions, and were not included in the perception test. For the naturalness rating, listeners were asked to judge whether they have heard a human utterance or a computer-generated sound.

The listening subjects were 20 native Beijing Mandarin speakers, who performed the perception tasks on Gorilla, an online experiment platform. They had no history of neurological or communication disorders, and passed a hearing screening at 20 dB HL bilaterally at 125, 250, 500, 750, 1000, 2000, 3000, and 4000 Hz.

3. RESULTS

3.1. Numerical evaluations

Table 2: Numerical assessments of tone learning separated by learning methods.

Learning method	RMSE	Correlation	Recog. rate
F_0 fitting	1.64	0.81	85%
Recog. only	4.12	0.18	66%
Recog.+deltaf_0	2.09	0.72	95%

Table 2 shows the root mean square error (RMSE), Pearson's correlation coefficient (*r*) and tone recognition rate for the three learning methods. As expected, learning by *f_0 fitting* worked well, achieving low RMSE and high correlation, consistent with previous findings on the same corpus [31]. Interestingly, the tone recognition rate, at 0.85, was also fairly high, which was consistent with [7]. *Recognition only* showed poor results, with high RMSE and low correlation, whereas *recognition+deltaf*$_0$ had the highest recognition rate at 95%, although its RMSE was higher and correlation was lower than those of f_0 fitting.

Figure 1 shows learning progression in terms of mean RMSE and recognition rate. Both indicators were recorded during learning with all methods, regardless of whether the method itself used them as optimization objectives. As can be seen, in all cases, RMSE was reduced over the iterations while recognition was improved. In all cases, the sudden improvement from iteration 450 was due to the shift from coarse- to fine-tuning phase of learning, as explained in 2.1.

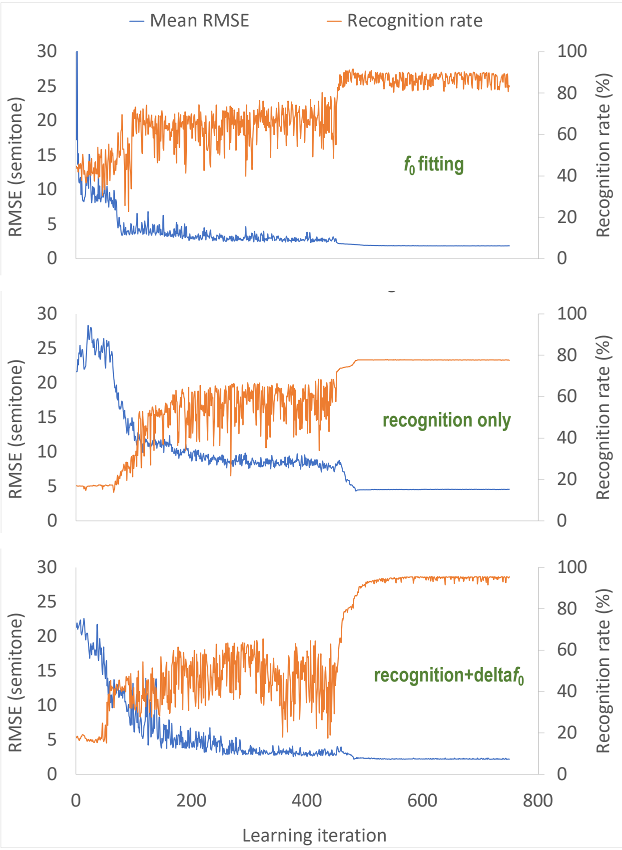

Figure 1: Examples of learning progression per iteration in terms of mean RMSE and recognition rate.

As can be seen, for f_0 fitting, RMSE was reduced to a very low level in the fine-tuning phase, but the increase of tone recognition hovered around 90%. For recognition only, both RMSE and recognition failed to improve much further in the fine-tuning phase. For recognition+deltaf_0, RMSE stopped to reduce below two semitones, but recognition went quickly above 95% after the onset of fine-tuning.

3.2. Human perceptual evaluation

Figure 2 shows perceptual tone identification rates for the four types of stimuli: a) original utterances, b) audios resynthesized with parameters learned from f_0 fitting, c) audios resynthesized with parameters learned from recognition only, and d) audios resynthesized with parameters learned from recognition+deltaf_0.

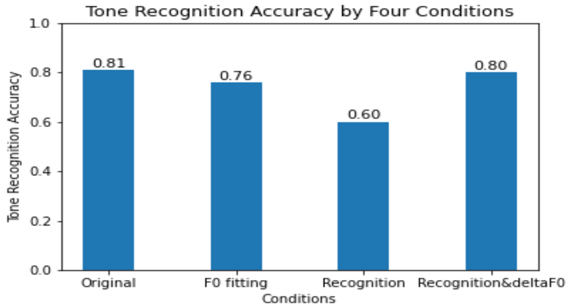

Figure 2: Tone recognition rate in four conditions.

The original utterances achieved the best tone recognition rate at 81%, while recognition+deltaf_0 was the second best at 80%. A two-tailed t-test found the difference between the two conditions non-significant. However, neither of these conditions performed nearly as well as the overall automatic tone recognition rate of 94.7% mentioned in section 2.3. One likely reason is that the recognizer performance was for tones of all syllables in each sentence. Syllables 1 and 5 both always had T1, the high-level tone, whose recognition rate was very high, partially due to over-training. For the tone of the second syllable alone, the recognizer achieved only 90% for the testing set, although this is still much higher than the 81% of the listener recognition of the original tones in Figure 2. Therefore, the superior performance of the recognizer is more likely because it has been trained on the same corpus, whereas the listeners relied on their real-life listening experience, which would include many more speakers with diverse tone articulations. The recognition rate for f_0 fitting was 76%, which was significantly lower than both the original ($t(16) = 6.57$, $p < 0.001$) and recognition+deltaf_0 ($t(16) = 4.89$, $p < 0.001$) conditions.

Figure 3: Heat map of confusion matrices for the perception of tones in four conditions.

Confusion matrices of tone perception are shown in Figure 3. These confusions can be examined together with the mean f_0 contours in Figure 4, which are separated by tone of the second syllable. First, for both T1 and T2, the recognition-only condition learned obviously wrong targets, high-fall for T1 and high-level for T2. Curiously, at 82% and 89%, the perception of these two tones did not seem to be severely affected. Second, For T2, the recognition+deltaf_0 condition generated a contour with an extra low minimum f_0 and a sharp terminal rise. This allowed the tone to be perceived (90%) as well as in the original condition (89%). Third, for T4, f_0 fitting generated a contour with a lower f_0 peak than both the original and recognition+deltaf_0 conditions. This is probably why it had a 26% confusion with T3. Finally, in the recognition+deltaf_0 condition, T3 was heard as T2 around 22% of the time. Although this is similar to the original condition where confusion with T2 was 12%, the f_0 contour in the bottom left plot of Figure 4 shows that the greater confusion was likely due to an earlier rise than the original T3.

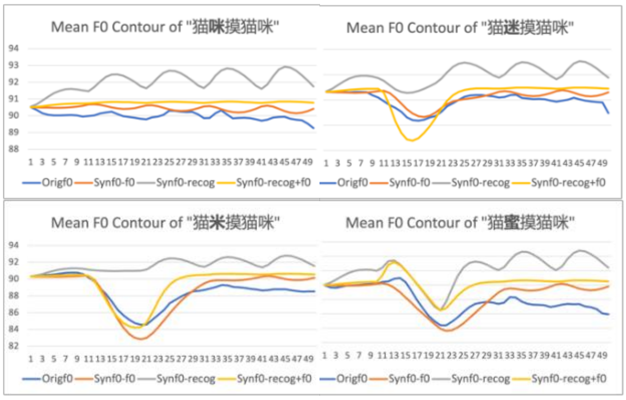

Figure 4: Mean F_0 contours, separated by the tone of the second syllable, clockwise from the top left: T1, T2, T3 and T4. The horizontal axis is normalized time (10 points/syllable). The vertical axis is f_0 in semitones.

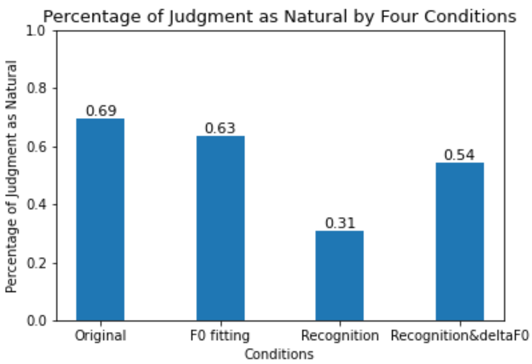

Figure 5: Percentage of utterances judged as natural speech (rather than synthetic) in the four conditions.

Figure 5 shows the results of naturalness judgment by listeners. As can be seen, even the original utterances were judged only 69% as human articulation. Interestingly, utterances from the f_0 fitting condition were judged as more likely to be humanly articulated than those from the recognition only ($t(16) = 7.55$, $p < 0.001$) and recognition+deltaf_0 ($t(16) = 4.97$, $p < 0.001$) conditions.

4. DISCUSSION AND CONCLUSION

This preliminary simulation study has demonstrated that perception-guided vocal learning [15, 26, 34] may also work for tone acquisition, provided that the target exploration range is constrained in the early learning phase, as done in the recognition+deltaf_0 condition. The quality of the learned tones with that method was better than those learned with f_0 fitting as assessed by both automatic tone recognition and human tone perception, except in terms of naturalness.

It is important to note, however, that the f_0 fitting method in this study was not strictly simulating direct mimicry, because it optimized f_0 contour match in all instances of each tone across all the repetitions by all speakers in the training set of the corpus (120 in total). But the optimization in recognition+deltaf_0 was also performed across all the utterances in the training set. In other words, the only difference between these two methods was the learning objective: to maximize the similarity of f_0 contours, or to maximize the tone recognition accuracy. On the face of it, the differences may be hard to comprehend. Why wouldn't achieving maximum acoustic similarity to multiple speakers lead to the best learning outcome? But the results clearly show an advantage for recognition-guided learning. This suggests that the difference in learning objectives is not trivial, as it may reflect *the core nature of speech as a communication system*. Given this nature, the proper objective of vocal learning should be to gain the ability to produce maximally intelligible speech rather than to just sound like other speakers. And the same may also be true of adult speech. That is, what makes a contrastive phonetic unit equivalent across different speakers is that it has been learned in such a way that it is most likely to be perceived as that unit. While this may sound circular as a factual definition, the circularity would disappear once it is treated as an operational definition, as shown in this study.

It is unclear, however, why recognition guidance alone did not work well in this study. Is it indeed due to a lack of cross-reference to other parameters as speculated in the Introduction? If yes, is it possible to introduce some minor adjustments to the current learning algorithm? This will need to be addressed in future studies.

5. REFERENCES

[1] Asada, M. (2016). Modeling early vocal development through Infant – Caregiver interaction: A review. *IEEE Transactions on Cognitive and Developmental Systems, 8* (2), 128-138.

[2] Beecher, M. D., & Brenowitz, E. A. (2005). Functional aspects of song learning in songbirds. *Trends in Ecology & Evolution, 20* (3), 143e149.

[3] Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International* **5:9/10**: 341-345.

[4] Brainard, M. S., & Doupe, A. J. (2002). What songbirds teach us about learning. *Nature, 417* (6886), 351-358.

[5] Brass, M. and Heyes, C. (2005). Imitation: is cognitive neuroscience solving the correspondence problem? *Trends in cognitive sciences* **9**(10): 489-495.

[6] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011.

[7] Chen, Y., Gao, Y. and Xu, Y. (2022). Computational modelling of tone perception based on direct processing of f0 contours. *Brain Sciences 12* (3), 337.

[8] Chen, Y. and Xu, Y. (2021). Parallel Recognition of Mandarin Tones and Focus from continuous F0. *1st International Conference on Tone and Intonation (TAI)*. Sonderborg, Denmark.

[9] Dehaene-Lambertz, G., Dehaene, S., & Hertz-Pannier, L. (2002). Functional neuroimaging of speech perception in infants. *Science, 298* (5600), 2013-2015.

[10] Fitch, W. T. & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America, 106* (31), 1511–1522.

[11] Howard, I. S. and Huckvale, M. A. (2005). Training a vocal tract synthesizer to imitate speech using distal supervised learning. In *Proceedings of International Conference on Speech and Computer (SPECOM)*, Patras, Greece**:** 159-162.

[12] Imada, T., Zhang, Y., Cheour, M., Taulu, S., Ahonen, A., & Kuhl, P. K. (2006). Infant speech perception activates Broca's area: a developmental magnetoencephalography study. *Neuroreport, 17* (10), 957-962.

[13] Johnson, K. (2005). Speaker normalization in speech perception. In *The handbook of speech perception*. D. B. Pisoni and R. E. Remez pp**.** 145-176.

[14] Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* **220**(4598): 671-680.

[15] Krug, P., Birkholz, P., Gerazov, B., Niekerk, D. R. v., Xu, A. and Xu, Y. (2023). Artificial vocal learning guided by phoneme recognition and visual information. IEEE Transactions on Audio, Speech and Language Processing.

[16] Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences* **97**(22): 11850-11857.

[17] Kuhl, P. K. (2003). Human speech and birdsong: communication and the social brain. *Proceedings of the National Academy of Sciences, 100* (17), 9645-9646.

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

[19] Lin, J., Li, W., Gao, Y., Xie, Y., Chen, N. F., Siniscalchi, S. M., … Lee, C.-H. (2018). Improving Mandarin Tone Recognition Based on DNN by Combining Acoustic and Articulatory Features Using Extended Recognition Networks. *Journal of Signal Processing Systems, 90* (7), 1077–1087.

[20] Pagliarini, S., Leblois, A. and Hinaut, X. (2021). Vocal Imitation in Sensorimotor Learning Models: A Comparative Review. *IEEE Transactions on Cognitive and Developmental Systems* **13**(2): 326-342.

[21] Philippsen, A. K., Reinhart, R. F. and Wrede, B. (2014). Learning how to speak: Imitation-based refinement of syllable production in an articulatory-acoustic model. In *Proceedings of 4th International Conference on Development and Learning and on Epigenetic Robotics*. IEEE**:** 195-200.

[22] Prom-on, S., Xu, Y. and Thipakorn, B. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *Journal of the Acoustical Society of America* **125**: 405-424.

[23] Skinner, B.P. (1957). *Verbal Behavior*. New York: Appleton/Century Crofts.

[24] Speidel, G. E., & Nelson, K. E. (1989). A fresh look at imitation in language learning. *In The many faces of imitation in language learning* (pp. 1-21). Springer, New York, NY.

[25] Tourville, J. A. and Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes* **26**(7): 952-981.

[26] Van Niekerk, D. R., Xu, A., Gerazov, B., Krug, P. K., Birkholz, P., Halliday, L., Prom-on, S. and Xu, Y. (2023). Simulating vocal learning of spoken language: Beyond imitation. Speech Communication 147: 51-62.

[27] Vorperian, H. K., & Kent, R. D. (2007). Vowel acoustic space development in children: A synthesis of acoustic and anatomic data.

[28] Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics 27: 55-105.*

[29] Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication* **46**: 220-251.

[30] Xu, Y. (2019). Prosody, Tone and Intonation. *In The Routledge Handbook of Phonetics*. W. F. Katz and P. F. Assmann: Routledge, New York. pp. 314-356.

[31] Xu, Y. and Prom-on, S. (2014). Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. *Speech Communication* 57, 181-208.

[32] Xu, Y. and Sun X. (2002). Maximum speed of pitch change and how it may relate to speech. *Journal of the Acoustical Society of America, 111* (3): 1399-1413.

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

[34] Xu, Y., Xu, A., Niekerk, D. R. v., Gerazov, B., Birkholz, P., Krug, P. K., Prom-on, S. and Halliday, L. F. (2022). Evoc-Learn — High quality simulation of early vocal learning. In *Proceedings of Interspeech 2022.*