# Chapter 13

# Modeling tone and intonation by simulating learning

Yining Ding[a], Yue Chen[b] & Yi Xu[a]
[a]University College London, England [b]Nanjing University of Aeronautics and Astronautics, China

Computational modeling is one of the best ways to study tone and intonation. The effectiveness of modeling, however, depends on not only the quality of the basic model, but also how the model is trained. The training process is not unlike how human learners acquire the ability to produce tone and intonation. Traditional understanding of language acquisition is that it is achieved mainly through imitation, which, when simulated in modeling can be done through $F_0$ curve-fitting. Recent research, however, has suggested an alternative, namely, production learning guided by perceptual recognition. In this study we tested this strategy further by simulating the learning of Mandarin tone and focus using an automatic tone-focus recognizer to train PENTAtrainer—an articulatory-functional modeling tool. Results show that, the addition of focus increased the learning difficulty compared to a previous simulation of tone learning only. But the recognizer trained learning models achieved performance levels close to that of $F_0$ curve-fitting, indicating a potential for improvement in future modeling work.

## 1 Introduction

A striking fact in language development is that children naturally acquire fluent and intelligible speech without explicit articulatory instructions (Kuhl 2004, Xu et al. 2022). Even in the absence of imitation templates or corrective feedback, they are able to develop a prosodically well-formed and comprehensible speech system through everyday interaction (Kuhl et al. 2008). This raises a fundamental

Yining Ding, Yue Chen & Yi Xu. 2026. Modeling tone and intonation by simulating learning. In Plinio A. Barbosa, Nick Campbell, Dafydd Gibbon, Keikichi Hirose & Daniel Hirst (eds.), *Developments in the Modeling of Speech Prosody*, 289–317. Berlin: Language Science Press. DOI: ??

question for both linguistic theory and computational modeling: How can learners acquire multi-layered prosodic functions—such as tone and focus—through perceptual mechanisms alone?

Recent research suggests that perceptual feedback may act as an internal guiding force in speech production (Guenther et al. 2006, Purcell & Munhall 2006, Xu et al. 2024). If this view is correct, prosodic modeling should go beyond simply reproducing surface acoustic contours. It should also simulate how overlapping prosodic functions are gradually acquired, encoded, and recognized over time (Meng et al. 2023, Chen & Xu 2025). Based on this idea, the present study focuses on Mandarin Chinese and examines how perceptual feedback affects the learning of prosodic features.

### 1.1 The role and development of computational models for prosody

Prosody, including intonation, is one of the most important components of human speech (Cutler et al. 1997, Wagner & Watson 2010, Xu 2019). At the same time, it is also one of the most elusive aspects (Taylor 2009, Wagner & Watson 2010), and remains poorly mastered even in today's AI technology (Barakat et al. 2024). This is despite decades of research that can be dated back to a beginning not much later than the study of segmental aspect of speech (Palmer 1922, Pike 1945, Lehiste 1970). One way to advance our understanding of prosody is to develop computational models of prosody, an approach pioneered by Fujisaki's early works (Fujisaki & Nagashima 1969, Fujisaki 1983, 1992). An important function of computational modeling is that it allows the testing of theories and conceptual models of prosody, as has been done in some studies (Pierrehumbert 1981). Some of the computational models were not originally proposed for testing any existing theories, but they may be considered as constituting their own theories (Fujisaki 1983, 1992, Hirst 2011). In other words, regardless of whether a computational model is claimed to be theoretical or not, if proven effective in simulating intonation, it may indeed have captured something theoretically relevant.

One common feature of computational models of intonation (which differs from purely theoretical models) is that they all aim at fitting model-generated $F_0$ contours to human-produced $F_0$ contours. What they differ is only in how and how closely the fitting is achieved. The simplest fitting is achieved by reducing the number of points in the original $F_0$ trajectory and replacing the missing points with straight or curved lines as in the stylization function in Praat (Boersma & Weenink 2001) and many other modeling works ('t Hart et al. 1990, Pierrehumbert 1981). The second approach is polynomial fitting of various type (Andruski & Costello 2004, Grabe et al. 2007). The third approach is to develop

algorithms to approximate natural $F_0$ contours based on various assumptions about the basic components of tone and intonation, such as shape of $F_0$ peaks (Taylor 2000), articulatory commands (Fujisaki & Nagashima 1969, Fujisaki 1983, 1992), or prosodic functions (Bailly & Holm 2005, Prom-on et al. 2009).

An important trend in intonation modeling, especially in the approximation approaches, is toward models that are trainable through analysis-by-synthesis (Fujisaki & Nagashima 1969). In these approaches, model parameters are optimized by repeatedly using them to generate $F_0$ contours which are compared to natural contours until a close match is found (Bailly & Holm 2005, Hirst 2004, Prom-on et al. 2009, Xu & Prom-on 2014). Interestingly, this trend is in the same direction of artificial intelligence (AI), especially large language models (LLMs). An obvious advantage of this trend is that it follows the principle that the best way to understand a system is to re-create it through training. What is less obvious, but likely even more meaningful, is that training is a form of learning, and learning is what has formed the system with all its characteristics in the first place. Following this idea, the best way to understand intonation is to know how children have learned it in speech acquisition.

## 1.2 Recognition-guided production learning

In the field of language acquisition, a long-standing question is how children manage to acquire speech in the absence of explicit instructions. It has been suggested that vocal learning is achieved by auditory imitation (Kuhl 2000, Vihman & de Boysson-Bardies 1994). The challenge, however, is that speech sounds are highly variable both within and across speakers, making it hard for children to know how to imitate. Worse, because their vocal anatomy is very different from adults' (Lieberman et al. 1972, Fitch & Giedd 1999), children's imitation cannot closely match adult speech. This has led some researchers to argue that vocal learning by imitation is impossible (Messum & Howard 2015). But a well-established behavioral finding is that infant's acquisition of speech perception precedes that of speech production (Werker & Tees 1984, Kuhl 2004). By 6 months of age, they have already started to distinguish between various language-specific phonetic categories (Werker & Tees 1984, Kuhl 2004), but they won't start to say their first words until 12 months. This means that they may have used their learned perception to guide their production learning (Kuhl et al. 2008), which is supported by the finding that speech discrimination skill in 6-month-old infants can predict later performance in speech production (Tsao et al. 2004). What remains unclear, however, is how exactly perception guides

the learning of articulation. A long-standing suggestion is that during perceptual learning, children store some kind of phonetic representations in the form of prototypes, exemplars or auditory patterns, which they then use as imitation goals during production learning (Kuhl 2007, Kuhl et al. 2008).

Recent computational simulations, however, have suggested an alternative that is even more plausible. That is, vocal learning could be done by trying to improve the accuracy of production assessed by children's own perceptual recognition of phonetic categories, without imitating exemplars or auditory patterns that are either live or stored in memory. This possibility first emerged in the Evoc-learn project (Xu et al. 2022), and has since been demonstrated to be effective in a series of studies on the simulation of vocal learning of CV syllables (Krug et al. 2023, Van Niekerk et al. 2023, Xu et al. 2024) and lexical tones (Meng et al. 2023). These simulations use analysis-by-synthesis as the basic learning mechanism (Ohman & Lindqvist 1965), but the optimization objective is to improve either the identification of the synthetic sound categories by a speech recognizer, or the acoustic similarity of the synthetic sounds to exemplar human utterances. The results have inevitably showed that recognition-guided production learning generated better outcome than imitation-oriented learning.

The effective simulation of recognition-guided tone learning raises the further question as to whether recognition-guided vocal learning also works for the learning of multiple prosodic functions, e.g., tone and focus, that are produced in parallel (Xu 2005). The present study is conducted to address this question.

### 1.3 Tone and focus in Mandarin Chinese

Mandarin Chinese is a prototypical tone language which uses fundamental frequency ($F_0$ ) to distinguish words in addition to spectral patterns related to consonants and vowels (Chao 1968). In addition to tone, however, Mandarin also uses $F_0$ to encode non-lexical functions such as focus (Xu 1999). The four tones of Mandarin have been shown to have underlying targets of high-level—Tone 1, rising—Tone 2, low—Tone 3 and falling—Tone 4 (Xu 1997, 1999). Focus is a pragmatic function for highlighting part of a sentence that the speaker wants to emphasize (Xu 1999). It is encoded by on-focus expansion of pitch range, syllable duration and intensity, and post-focus compression of pitch range and intensity. There is no consistent variation for pre-focus words, however. Of the three acoustic dimensions, pitch is by far the most important, because its manipulation alone by computational modeling is sufficient to allow listeners to recognize focus (Prom-on et al. 2009, Xu & Prom-on 2014), and duration manipulation adds only small enhancement (Prom-on et al. 2009).

## 1.4 Simulating recognition-guided production learning of tone and focus

Inspired by the Evoc-learn project (Xu et al. 2022), Meng et al. (2023) investigated whether recognition-guided production learning would also be effective in simulating the learning of Mandarin tone production. Unlike Xu et al. (2024) which simulated the recognition-guided learning with a recognizer trained on the LibriSpeech corpus (Panayotov et al. 2015), but simulated the imitation-guided learning by acoustic matching of specific natural utterances, Meng et al. (2023) used the same tone corpus (Xu 1999) to train both the tone recognizer and the tone production learning processes. This allowed them to directly compare three different learning strategies:

1. $F_0$ fitting—minimizing root-mean-square error (RMSE) between synthetic and original $F_0$ contours,
2. recognition-only—optimizing tone classification accuracy by a pretrained recognizer based on support vector machine (SVM), and
3. recognition + delta$F_0$—jointly minimizing recognition error and distance of mean $F_0$ between synthetic and corresponding original utterances.

The trainings were done with a revised version of PENTAtrainer (Xu & Prom-on 2014), and the output was evaluated by the SVM tone recognizer as well as human listeners. The recognition + delta$F_0$ strategy achieved the highest tone recognition rate (95% SVM recognition, 80% human recognition). Although $F_0$ fitting yielded the lowest RMSE, it performed worse in both automatic and human tone recognition.

## 1.5 Learning the production of tone and focus in parallel

Like Xu et al. (2024), Meng et al. (2023) simulated the learning of only a single communicative function, namely, lexical contrast, through CV syllables in the former, and syllabic tones in the latter. But speech signals convey meanings not just through lexical contrasts, but also through other functions, as characterized by the parallel encoding and target approximation (PENTA) model shown in Figure 1.

As depicted, the continuous surface dynamic acoustic signals are directly generated by a sequential articulatory Target Approximation (TA) process (Xu & Wang 2001, Xu 2005). But the parameters, *height* (or *position*), *slope*, and *strength*,

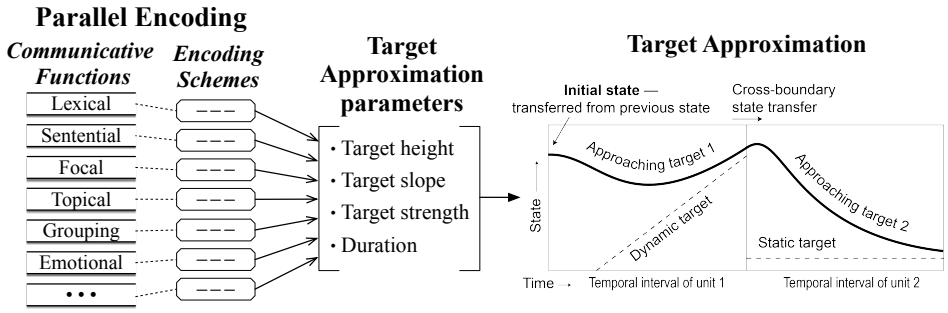



Figure 1: Schematic representation of the PENTA model (Xu 2005). The dashed lines in the right-most section represent the underlying targets specified by the target approximation (TA) parameters in the middle section (Xu & Wang 2001, Prom-on et al. 2009). These parameters are jointly determined by multiple communicative functions (left-most section) through their respective encoding schemes. Theses parameters ultimately control the TA process that generate the surface dynamic acoustic signals represented by the solid trajectories.

that control the articulation process are jointly determined by multiple communicative functions through their respective encoding schemes. Such an encoding process makes it hard for learners to identify the communicative functions either directly from surface acoustic signals or indirectly through the underlying targets. To resolve this difficulty for perception, Chen & Xu (2025) tested the possibility that the perception of the tone and focus categories can be achieved by simply training a recognizer with syllable-sized continuous $F_0$ contours that are functionally labeled. They tested this idea again on the Xu (1999) corpus, but this time with all the four focus conditions included (only the neutral-focus sentences were modeled in Meng et al. (2023)). The results showed that both tone and focus categories can be recognized through training with high accuracy, and this was achieved whether the training was on the classification of tone and focus as conjoined categories, or as fully independent functions. The training did not include an intermediate goal to recognize the underlying pitch targets suggests, however. In other words, not only does perceptual recognition of functional categories do not require the development of knowledge of the articulation process as suggested by motor theory (Liberman & Mattingly 1985), but also the underlying association of surface acoustics with multiple communicative functions can be discovered simply by brute-force training on labeled syllable-sized trajectories. The power of such articulatory-knowledge-free perception learning then raises the possibility that the learned recognition system can guide the learning of the articulatory production of the multiple functions. The current study is conducted

to test this possibility by simulating recognition-guided production learning of tone and focus.

# 2 Method

## 2.1 Speech corpus

The speech corpus used in this study was adapted from the experimental dataset initially established by Xu (1999) for investigating tone and prosodic focus in Mandarin.

The full corpus consists of recordings from eight native Mandarin speakers (four females and four males), all producing fluent and natural speech at an average rate of approximately five syllables per second. Each utterance consists of two disyllabic words and one monosyllabic word, with tonal variations on the second, third and fourth syllables, as shown in Table 1.

Table 1: Sentences used as recording materials and their tone patterns. H, R, L, and F represent high, rising, low, and falling tones, respectively (Xu, 1999).

| **Word1** | **Word2** | **Word3** |
|---|---|---|
| HH 猫咪 /māomī/ 'Kitty' | H 摸 /mō/ 'touches' | HH 猫咪 /māomī/ 'kitty' |
| HR 猫迷 /māomí/ 'Cat-fan' | R 拿 /ná/ 'takes' | LH 马刀 /mǎdāo/ 'sabre' |
| HL 猫米 /māomǐ/ 'Cat-rice' | F 卖 /mài/ 'sells' | |
| HF 猫蜜 /māomì/ 'Cat-honey' | | |

Each of these base sentences was spoken under four focus conditions: focus on the first word (initial), second word (medial), third word (final), and neutral focus, elicited by WH-questions. This resulted in a total of 3840 utterances (24 base sentences × 4 focus conditions × 8 speakers × 5 repetitions).

For training and evaluation purposes, recordings from six speakers (three females and three males) were used as the training set, totaling 2880 utterances. Recordings from the remaining two speakers (one female and one male), totaling 960 utterances, were reserved as the test set.

## 2.2 Modeling Framework: The extended PENTAtrainer

### 2.2.1 The original PENTAtrainer

The PENTA model (Xu 2005) presented in 1.5 was first quantified in Prom-on et al. (2009) for the TA part of the model:

$$c_1 = f_0(0) - b \quad (1)$$

$$c_2 = f_0'(0) + c_1\lambda - m \quad (2)$$

$$c_2 = (f_0''(0) + 2c_2\lambda - c_1\lambda^2)/2 \quad (3)$$

Where $F_0$ represents surface $F_0$ contours, $b$, $m$ and $\lambda$ represent the height, slope and strength of the pitch targets, respectively, and $c_1$, $c_2$ and $c_3$ are transient coefficients determined jointly by the target and the initial $F_0$ of the target interval.

PENTA was then made fully trainable in the form of PENTAtrainer to allow automatic simulation of the parallel encoding aspect of the model (Prom-on & Xu 2012, 2013, Xu & Prom-on 2014). PENTAtrainer operationalizes the learning of the multi-functional pitch targets through $F_0$ fitting, i.e., minimizing the difference between TA-generated and original $F_0$ contours measured in root mean square error (RMSE).

### 2.2.2 Recognition-driven PENTAtrainer

To simulate recognition-guided learning, Meng et al. (2023) modified the original PENTAtrainer to include tone recognition error as an alternative loss to RMSE. The recognition error was generated by a tone recognizer that was externally implemented in Python using the scikit-learn library. It was a support vector machine (SVM) trained on syllable-sized $F_0$ contours. The model was capable of recognizing tone with high accuracy (Chen et al. 2022). For the present study, two new recognizers were developed based on SVM and multilayer perceptrons (MLP), respectively. They were both trained on the full corpus of Xu (1999) with all the four focus conditions as did in Chen & Xu (2025).

To test the effectiveness of recognition-guided tone and focus learning, three learning strategies were implemented:

$F_0$ fitting—The baseline method: minimizing RMSE between generated and original $F_0$ contours. This strategy largely simulates imitation-based learning.

Recognition only—Learning by only minimizing recognition error generated by an automatic tone-focus recognizer. This strategy simulates learning guided by perceptual recognition of phonetic categories.

Recognition+delta$F_0$—Learning by minimizing both recognition error and the utterance-wide difference in mean $F_0$ (delta $F_0$). This strategy simulates learning

guided by a combination of perceptual recognition and imitation. The combined objective function is defined as:

$$e = 0.9e_r + 0.1d \tag{4}$$

where $e_r$ = 10 (1 − recognition rate [0,1], and $d = f_{0,\text{orig}} - f_{0,\text{orig}}$) (Meng et al. 2023). Unlike Meng et al. (2023), however, delta$F_0$ was applied through the entire training process rather than only in the global-tuning phase.

## 2.3 Recognition models

### 2.3.1 Support vector machine (SVM)

Support Vector Machines (SVMs), which are widely used in classification tasks, aim to find the optimal hyperplane $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ that maximizes the margin between classes. This maximum-margin principle offers notable advantages for classifying pitch patterns by supporting strong generalization ability and robustness to overfitting. Figure 2 gives a simple example of a binary linear SVM classifier.

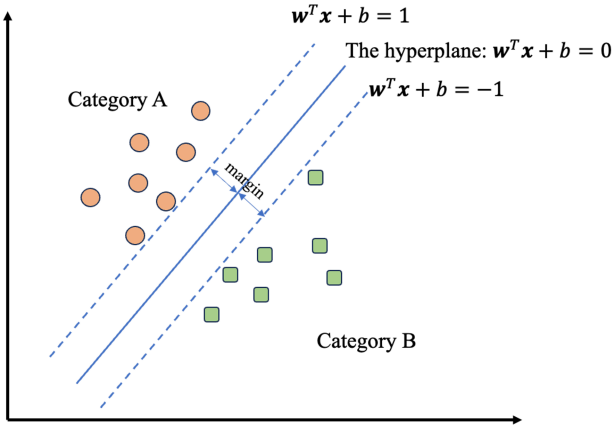



Figure 2: A simple demonstration of a linear SVM.

This formulation balances the goals of margin maximization and classification error minimization. Since $F_0$ features, such as time-normalized pitch contours encoding tonal shape, slope, and curvature, are often not linearly separable, we employ the radial basis function (RBF) kernel to transform the input data into

higher-dimensional space to enable linear separation of complex prosodic patterns.

Following Chen & Xu (2025), we adopted an RBF-kernel SVM for the tone and focus recognition task in this study. We formulated the classification as a multi-class task using a one-versus-one (OVO) strategy, which splits the $n$-class classification task into $n(n-1)/2$ binary tasks, and resolved predictions via a voting strategy (Kreßel et al. 1999). We trained the soft-margin SVM using the SVC module in scikit-learn (Pedregosa et al., 2011), with hyperparameters C and γ selected via a grid search using five-fold cross-validation. To ensure robustness against variation from data partitioning and random initialization, each hyperparameter configuration was evaluated over five independent runs. The final model selection was based on the average cross-validation performance. The fully trained SVM recognizer achieved 89.18% tone accuracy and 71.31% focus accuracy for the full corpus.

### 2.3.2 Multilayer perceptrons (MLPs)

Multilayer Perceptrons (MLPs) are powerful supervised learning models capable of capturing complex nonlinear relationships between input features and output labels (Cybenko 1989).

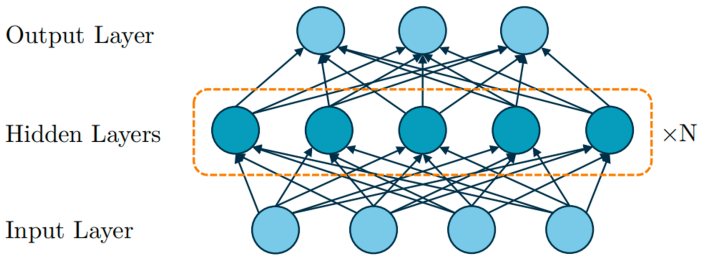



Figure 3: A schematic illustration of a Multilayer Perceptron (MLP).

As illustrated in Figure 3, an MLP could transform input $F_0$ vectors through multiple layers of affine transformations followed by non-linear activations. This layered architecture enables the model to learn implicit features of pitch-related information, thereby capturing latent prosodic structures without relying on handcrafted acoustic similarity metrics or rule-based systems. Empirical evidence supports the effectiveness of MLPs in pitch-related tasks. For example, Drugman et al. (2018) reported a 45% relative reduction in voicing detection errors compared to traditional pitch extraction methods, highlighting the superiority of neural architectures for acoustic modeling. Building on previous neural

simulations of tone and focus perception (Chen & Xu 2021, 2025), we trained MLPs with Pytorch (Paszke et al., 2017) taking time-normalized $F_0$ contours extracted from annotated Mandarin utterances as input features. The whole dataset was divided into training and validation set with a ratio of 3:1. Model performance was assessed with five independent runs. The fully trained MLP recognizer achieved 87.52% tone accuracy and 63.20% focus accuracy for the full corpus.

## 2.4 Learning target parameters through training

The three learning strategies, $F_0$ fitting, recognition-only, and recognition + delta$F_0$, were implemented in five training configurations as shown in Table 2. Each configuration was run five times to even out variation due to random initialization.

Table 2: Learning strategies simulated by five training configurations.

**Table 2.** Overview of learning strategies and recognizer configurations.

| Strategy Type | Recognizer Type | Description |
|---|---|---|
| $F_0$ fitting | None | Baseline |
| Recognition-only | SVM | Recognition-guided learning with probability feedback |
| Recognition-only | MLP | Recognition -guided learning with probability feedback |
| Recognition+delta$f_0$ | SVM | Recognition -guided learning with probability feedback |
| Recognition+delta$f_0$ | MLP | Recognition -guided learning with probability feedback |

Each training process consisted of 700 iterations, which were divided into two phases. The first was the global-tuning phase consisting of 450 iterations, during which parameters of all pitch targets were optimized at once. The second was the fine-tuning phase consisting of the rest of the 250 iterations, during which parameters of each specific tone-focus syllable were optimized one set at a time.

## 2.5 Automatic evaluation of learning outcome

Upon completion of the trainings, the learned target parameters were averaged across the five runs of each training configuration to obtain five parameter sets. PENTAtrainer was then run again to resynthesize all the utterances in the testing set (with utterances by one female and one male speakers) with the model-generated $F_0$ contours. The utterances with synthetic $F_0$ contours from all the training configurations were first assessed by the SVM-based recognizer for accuracy of tone and focus identification. They were then assessed by native Mandarin listeners in an online perception experiment.

The stimuli consisted of resynthesized utterances generated by parameters learned through the five training configurations (sound sources 2-6), along with

the corresponding original utterances (sound source 1). A total of 96 unique sentence combinations were used, based on a full factorial design with 4 initial words (w1) × 3 medial words (w2) × 2 final words (w3) × 4 focus conditions × 2 test speakers × 6 sound sources = 1152 sentences, as shown in Figure 4.

To control the duration of the experiment, only one of the five available repetitions of each sentence was selected at random. All stimuli were evenly distributed across participants, with each sentence evaluated five times. Each participant was assigned to judge 288 sentences, ensuring a balanced exposure and an acceptable duration (under 60 minutes).

The online tests were done through the Gorilla Experiment Builder platform (www.gorilla.sc), with participant recruitment done via Prolific (www.prolific.co). A total of 20 native speakers of Mandarin Chinese (10 females and 10 males), all aged 18 or above, participated in the study. None reported hearing impairments. Before each test session, the subject signed an online consent form and confirmed that they were in a quiet environment suitable for the listening task.

Each trial consisted of a fixed sequence of three-part perceptual judgment task. Upon hearing a five-syllable utterance, the participant was asked to complete the following:

- Tone Judgment: Identifying the lexical tone of the second syllable by selecting the corresponding choice displayed in Chinese, as shown in Fig. 4. Syllables in other positions were not included because they were not in minimal contrasts as the syllable structures varied with the tones.

- Focus Judgment: Identifying the location of focus in the sentence.

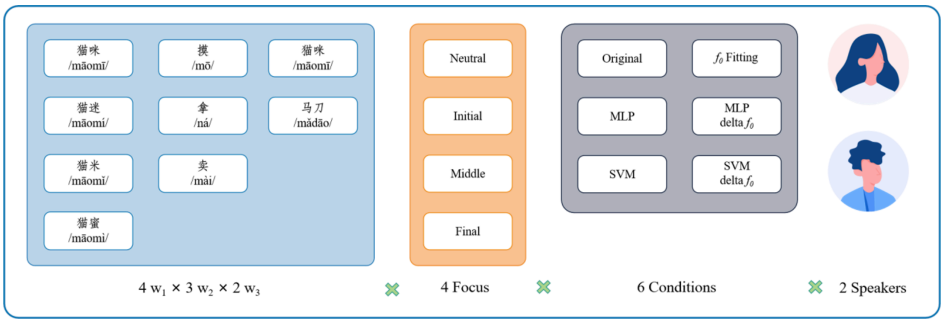


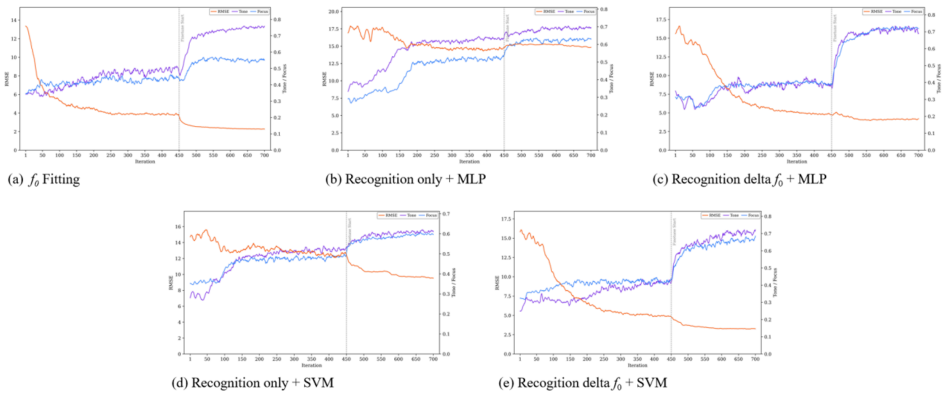

Figure 4: Experimental design showing the full factorial combination of lexical items, focus conditions, sound sources (conditions), and speakers.




Figure 5: Perception test interface with instructions and choices with selection buttons.

- Naturalness Judgment: Judging whether the utterance sounded natural or synthetic.

Before the formal experiment, all participants received a set of instructions and a list of possible sentence combinations, which helped to reduce confusion about lexical content that could affect tone identification. They were then given three practice trials (without feedback) to get familiar with the task interface and procedure. In each trial, only a single playback of the audio was allowed. Upon completion of all the choice selection, the next trial began automatically.



(a) $f_0$ Fitting (b) Recognition only + MLP (c) Recognition delta $f_0$ + MLP

(d) Recognition only + SVM (e) Recogition delta $f_0$ + SVM

Figure 6: Examples of learning progression per iteration in terms of mean RMSE and recognition rate.

# 3 Results

## 3.1 Model Training Performance

Figure 6 shows examples of learning progression in terms of RMSE and recognition accuracy of tone and focus across all 700 training iterations. Under the $F_0$ fitting configuration, RMSE decreased sharply and eventually stabilized to a low level in the fine-tuning phase. This strategy did not use tone or focus recognition as the optimization objective, but steady gains in tone recognition accuracy was still achieved.

Among the two recognition-only configurations, the SVM model achieved an accuracy of 62.49%, clearly outperforming the MLP model at 50.48%. With the additional delta$F_0$ loss, the MLP model showed a notable improvement to 57.85%,

Table 3: Automatic evaluation on the test set: mean RMSE and correlation of $F_0$ trajectories.

| Method | Mean RMSE | Mean Correlation |
|---|---|---|
| $F_0$ Fitting | 2.254 | 0.792 |
| SVM Recognition only | 9.480 | 0.212 |
| MLP Recognition only | 14.778 | 0.282 |
| SVM Recognition + delta$f_0$ | 3.231 | 0.653 |
| MLP Recognition + delta$f_0$ | 4.053 | 0.626 |

while the SVM-delta$F_0$ model showed a slight increase to 63.01%. But it is still much lower than that of the original speech at 87.92%.

In both of the recognition-only configurations, recognition accuracy increased rapidly during the global-tuning phase, but RMSE remained consistently high throughout, with no significant reduction in the fine-tuning phase.

When delta$F_0$ was introduced as an additional loss, model behavior changed markedly. For both of the recognition+delta $F_0$ configurations, recognition accuracy improved sharply after the onset of fine-tuning, while RMSE continued to decline, albeit more gradually.

Table 3 shows root mean square error (RMSE) and Pearson correlation ($r^2$) of each training configuration. As can be seen, the $F_0$ fitting method showed stable performance, with a low RMSE of 4.25 st and the highest correlation of 0.615. Recognition+delta$F_0$ also performed well. SVM + delta$F_0$ achieved the lowest RMSE among recognition-guided models, with a value of 3.996 st. MLP + delta$F_0$ had a slightly higher RMSE of 4.597 st, but reached the highest correlation 0.557, second only to $F_0$ fitting. In contrast, the recognition-only configurations performed less well. SVM produced a relatively high RMSE of 5.104 st and the lowest correlation of 0.280. MLP performed worst overall, with the highest RMSE of 8.020 st and a low correlation of 0.388.

### 3.1.1 Recognizer assessed tone classification accuracy

The accuracy of tone classification is assessed by the SVM-based recognizer (the one used in the SVM training configurations) for all five training configurations, and the results are shown in Figure 7. Also included in the figure is the accuracy of tone classification for the original $F_0$ contours (the dashed horizontal line) assessed by the same SVM recognizer. The $F_0$ fitting configuration achieved the highest accuracy at 68.80%. But it is still much lower than that of the original speech at 87.92%.

### 3.1.2 Recognizer assessed focus classification accuracy

Figure 8 shows the accuracy of focus classification assessed by the SVM-based recognizer for all five training configurations. Overall, focus classification accuracy remained relatively low across all training configurations. $F_0$ fitting yielded the highest accuracy of 31.15%, which is well below that the original natural recordings at 70.32%, as indicated by the dashed line. Among the two recognition-only configurations, the SVM-based model achieved an accuracy of 27.26%, slightly outperforming the MLP-based model at 24.29%. When delta$F_0$

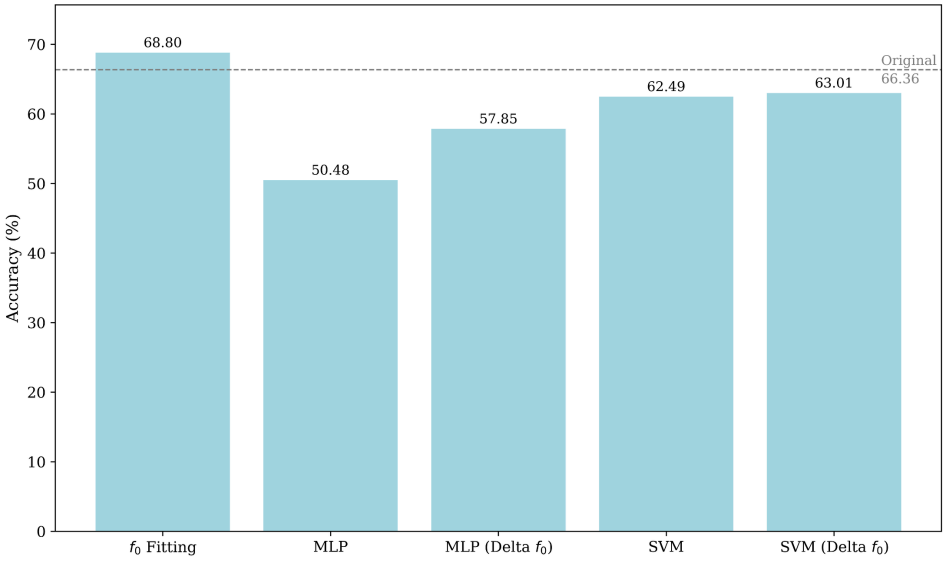



Figure 7: Automatic tone recognition accuracy across models.

was included as an additional loss, the MLP model improved to 28.91%, but the SVM model decreased to 26.73%.

## 3.2 Human perceptual evaluation

### 3.2.1 Tone perception

The ultimate test of the simulated learning of tone and focus articulation is human perception of the synthesized utterances. Figure 9 displays the average tone recognition accuracy across six sound sources. As expected, natural speech achieved the highest tone recognition accuracy at 80.10%, serving as a performance benchmark. Among synthesized stimuli, the $F_0$ fitting model reached 60.27%. The SVM model without delta$F_0$ loss also achieved a comparable result with 60.00%. In contrast, the MLP model yielded the lowest accuracy at 48.44%. The introduction of delta$F_0$ led to modest improvements for MLP (+7.5%) but a decline for SVM (–10.8%).

Figure 10 shows confusion matrices of tone of the second syllable (Tone 1–Tone 4) for all models. The original utterances and the $F_0$ fitting model both performed well, with high accuracy rates and strong diagonal weights, indicating clear tone distinctions across all four categories. In contrast, the MLP and SVM models, especially those without delta$F_0$ , show severe confusions, particularly for Tone

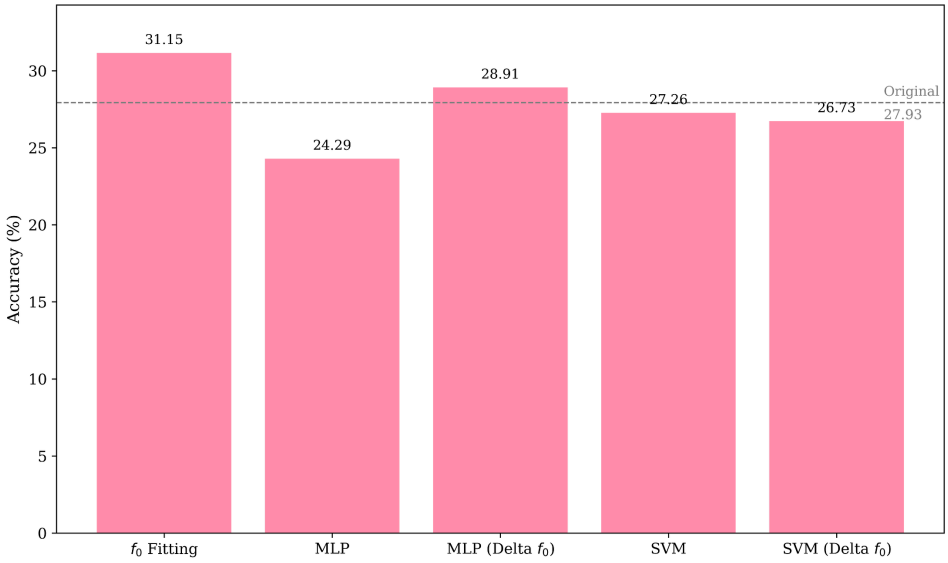



Figure 8: Automatic focus recognition accuracy across models.

3, which is frequently heard as Tone 2. Interestingly, the introduction of delta$F_0$ greatly improves the recognition of Tone 3 in both MLP (from 44% to 85%) and SVM (from 53% to 81%). However, this benefit came with a cost: the confusion between Tone 1 and Tone 2 increased dramatically. For example, in MLP, the rate of T1 being misclassified as T2 jumped from 12% to 38%.

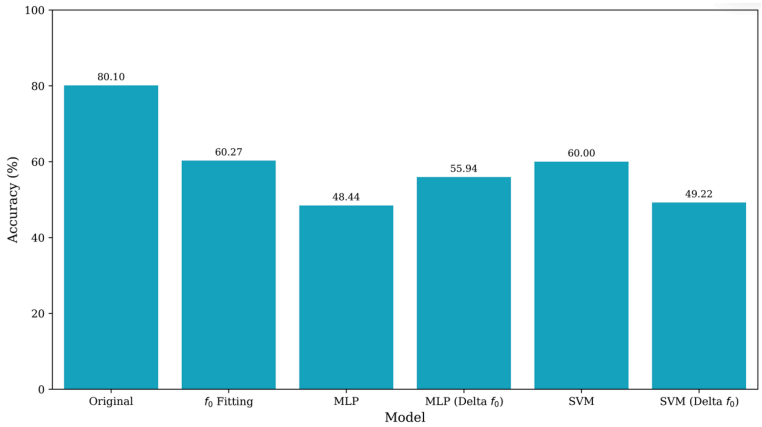



Figure 9: Human perception accuracy of tone recognition across models.

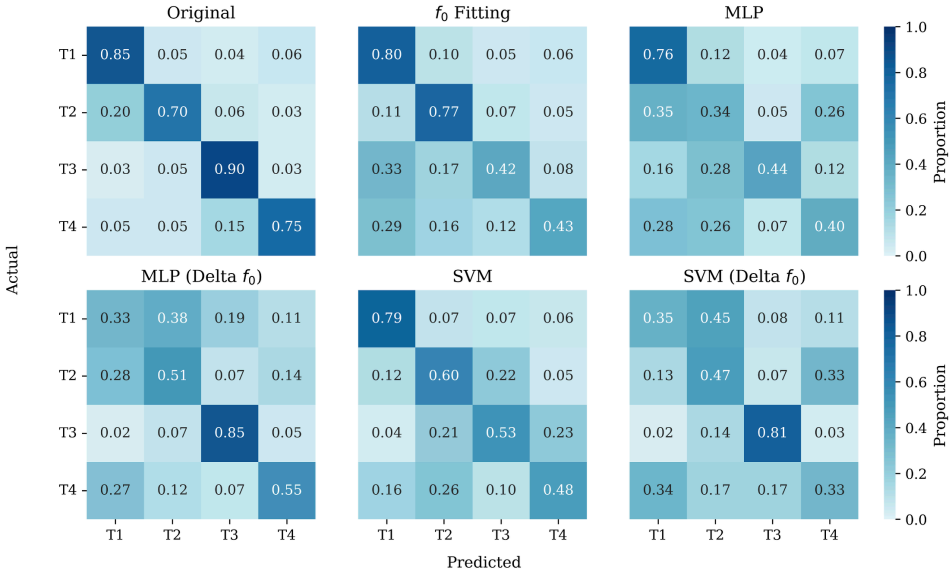



Figure 10: Heatmaps of confusion matrices of tone recognition in human perception across different models.

### 3.2.2 Focus perception

Figure 11 presents accuracies of focus identification for all models. All synthesized models except the unconstrained SVM achieved higher average accuracy than the original stimuli, which reached 54.27%.

Among the models, SVM with Delta$F_0$ loss yielded the highest accuracy at 59.11%, followed closely by MLP and MLP + delta$F_0$ with 58.13% and 58.75%, respectively. Both models outperformed $F_0$ fitting, which reached 57.46%. In contrast, the SVM model without any structural constraint showed the lowest performance with an accuracy of 42.60%.

Figure 11 shows that all models achieved the highest recognition accuracy on Medial focus, followed by Initial focus. In contrast, final focus is the most difficult to identify. This result is consistent with earlier studies, which found that sentence-final focus is often acoustically less salient than non-final focus due to prosodic boundary effects (Botinis et al. 1999, Xu et al. 2012). The overall trend observed in this study also supports findings from Yuan (2011) and Liu (2009), who reported that, in Mandarin declarative sentences, the recognizability of focus positions typically follows the order: medial > initial > final.

To further examine model performance across different types of prosodic prominence, we evaluated listener accuracy on four focus conditions: Neutral,

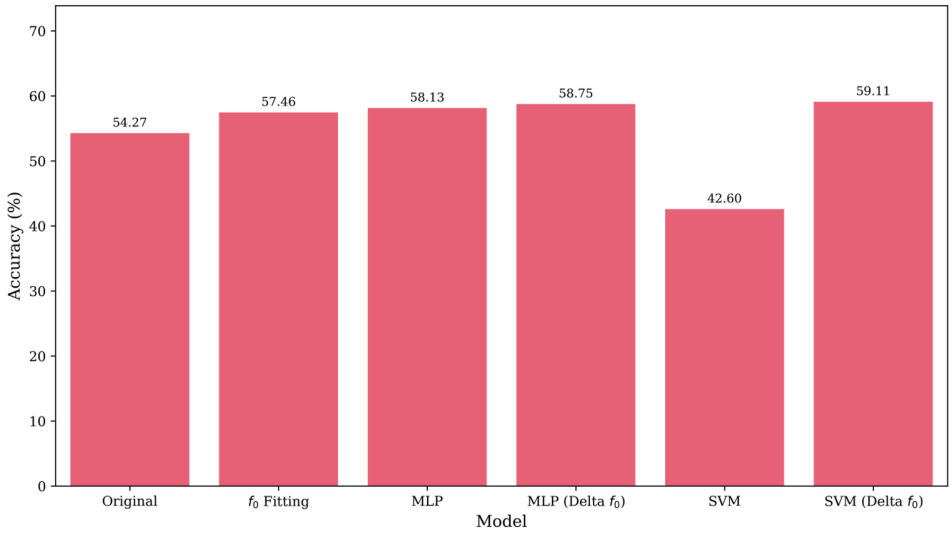



Figure 11: Human perception accuracy of focus recognition across models.

Initial, Medial, and Final.

A noteworthy pattern is that final focus is often misclassified as neutral focus, which is true of both the original utterances and the synthetic ones. This is consistent with previous many findings (Botinis et al. 1999, Rump & Collier 1996, Liu & Xu 2005, Chen & Xu 2025). This again provides support for the argument (Liu & Xu 2005, Xu 2019) against the notion of "broad focus" or "sentence-final default focus" (Büring 2006, Ladd 2008, Carlson et al. 2009, Yan & Calhoun 2020). Instead, it suggests again that neutral focus is the most likely default perceived focus, at least for Mandarin.

### 3.2.3 Naturalness Judgment

As stated in 2.5, the perceived naturalness was evaluated in a binary classification task. Participants judged whether each utterance sounded human-produced or machine-generated. The percentage of "human" responses was used as an indicator of perceived naturalness.

As shown in Figure 13, the original recordings received the highest naturalness score of 74.90%. The $F_0$ fitting configuration followed with a moderate score of 55.89%. In contrast, the MLP recognition only configuration showed much lower performance at 27.60%. And the SVM recognition only configuration performed better at 47.40%. Interestingly, adding the delta$F_0$ loss had mixed effects. For the

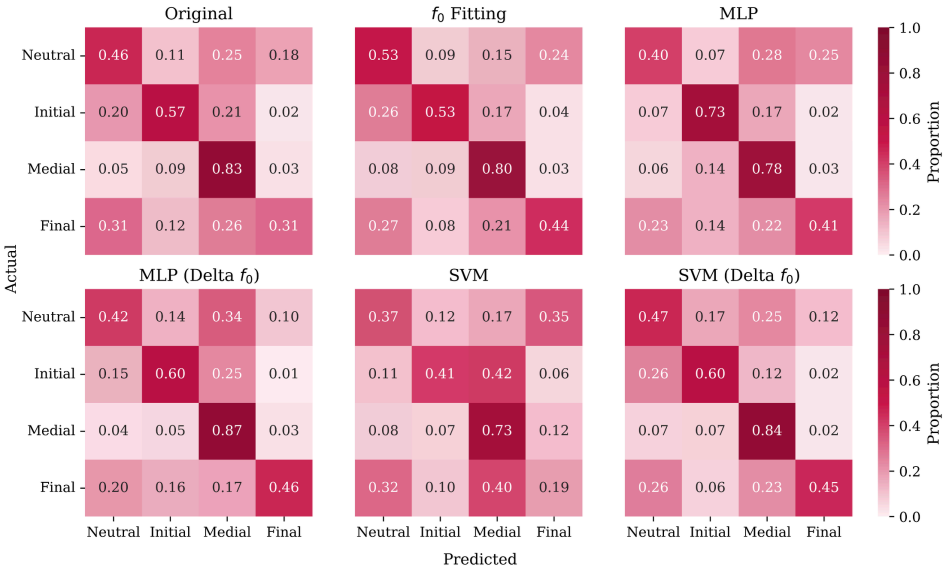



Figure 12: Heatmaps of confusion matrices of focus recognition in human perception across different models.

MLP model, it improved naturalness from 27.60% to 44.27%, whereas for the SVM model, it slightly reduced naturalness from 47.40% to 41.10%.

Perhaps the most interesting is that perceived naturalness in Figure 13 is highly consistent with tone perception accuracy in Figure 9. Across the speech of different sources, from original speech to that generated by different learning models, the greater the accuracy of tone perception, the more the speech is judged as natural. In contrast, the focus perception accuracies in Figure 11 show no resemblance to the perceived naturalness in Figure 13. Listeners, therefore, seem to be much more sensitive to the integrity of the syllabic tones than to that of sentential focus.

## 4 Discussion

The present study used computational simulation to examine the effectiveness of different strategies of learning tone and focus at the same time, using both an automatic tone-focus recognizer and human listeners to assess the learning performance. We trained multiple versions of tone and focus modeling systems, including $F_0$ fitting and recognition-based models with or without delta$F_0$

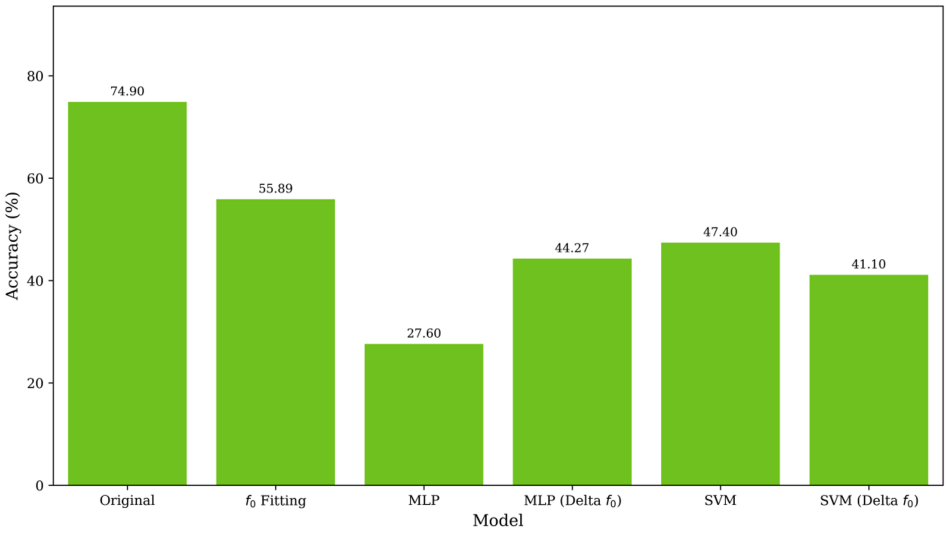



Figure 13: Percentage of utterances judged as natural speech (rather than synthetic) from the six sound sources.

loss. We evaluated all models using both automatic evaluations and human perception tests. Overall, the results show three main findings: (a) The $F_0$ fitting model demonstrated more consistent performance across most evaluation metrics, while recognition-guided learning yielded advantages in specific tasks, such as focus recognition by human listeners; (b) The inclusion of delta$F_0$ as a loss improved performance in some cases, but its effectiveness varied depending on model architecture and training objectives; (c) Automatic evaluations do not always align with human perception, especially in tasks involving focus recognition.

## 4.1 Evaluating recognition-guided learning versus $F_0$ fitting

As shown in Figure 10, in perception-based evaluation, recognition-guided learning models, particularly those combining recognition and delta$F_0$, show competitive performance in human focus recognition tasks. However, the $F_0$ fitting model shows more stable results across both automatic evaluations and the other two human perception tasks.

Interestingly, training configurations with delta$F_0$ as a loss did not consistently outperform $F_0$ fitting in the human perception evaluations. One possible explanation is that the inclusion of delta$F_0$ increased the demands on the model. It must

both classify prosodic categories accurately and generate $F_0$ contours with optimal $F_0$ height. If these objectives indeed partially conflict with each other, the delta$F_0$ constraint may introduce trade-offs that interfere with the learning of core prosodic patterns. The exact causes of these outcomes are of course unclear and should be explored further in future research.

### 4.2 Effect of delta$F_0$ across Different Model Architectures

We have observed a clear contrast in how delta$F_0$ affects learning outcomes across model architectures. For MLP, adding delta$F_0$ consistently improves performance in both automatic evaluations and human perception tests. In contrast, the results for SVM are mixed. On the one hand, adding delta$F_0$ greatly improves performance in focus recognition in human perception, with accuracy increasing from 42.60% to 59.11%, which is the highest score among all six sound sources. On the other hand, tone recognition under human perception shows a notable decline, dropping from 60.00% to 49.22% after delta$F_0$ was introduced.

Model architecture differences may explain the inconsistent effect of delta$F_0$ across MLP and SVM. The MLP, as a deep network with rich parameters and end-to-end training, is better suited for learning continuous and dynamic features. Prior studies have shown that neural networks can accurately model prosodic parameters such as $F_0$ contours, duration, and intensity in speech synthesis tasks, often outperforming traditional models in both accuracy and smoothness (Reddy & Rao 2016). In contrast, SVM is a static classifier whose decision boundary depends on fixed feature representations. Without mechanisms to handle sequential patterns, highly dynamic features like delta$F_0$ may be treated as noise or redundant information.

### 4.3 Discrepancies between human perception and automatic evaluation

A comparison of Figures 7 and 9 shows that there is much similarity between automatic evaluation and human perception of tone, with the only apparent exception of SVM+delta$F_0$, where the recognizer evaluation reached 63.01%, while human perception was only 49.22%. In contrast to tone recognition, focus identification showed much more discrepancy when comparing Figure 8 and Figure 11. While most models performed poorly in the automatic evaluations, their accuracies were substantially better in human perception. Most surprisingly, all the synthetic utterances, except those generated by the SVM only configuration, yielded better focus recognition than the original human utterances. A likely clue may

be found in what is known about the global nature of focus. It is encoded not only through local $F_0$ raising, but also through global pitch range modifications, including on-focus expansion and post-focus compression (Xu 1999, 2005). As can be seen in Figure 14, compared to the original $F_0$ contours, the synthetic ones show exaggerated on-focus raising as well as post-focus lowering of pitch ranges, especially in the case of the MLP model. These exaggerated patterns, though having an adverse effect on the automatic focus recognition, may have enhanced perception of focus by our listeners.

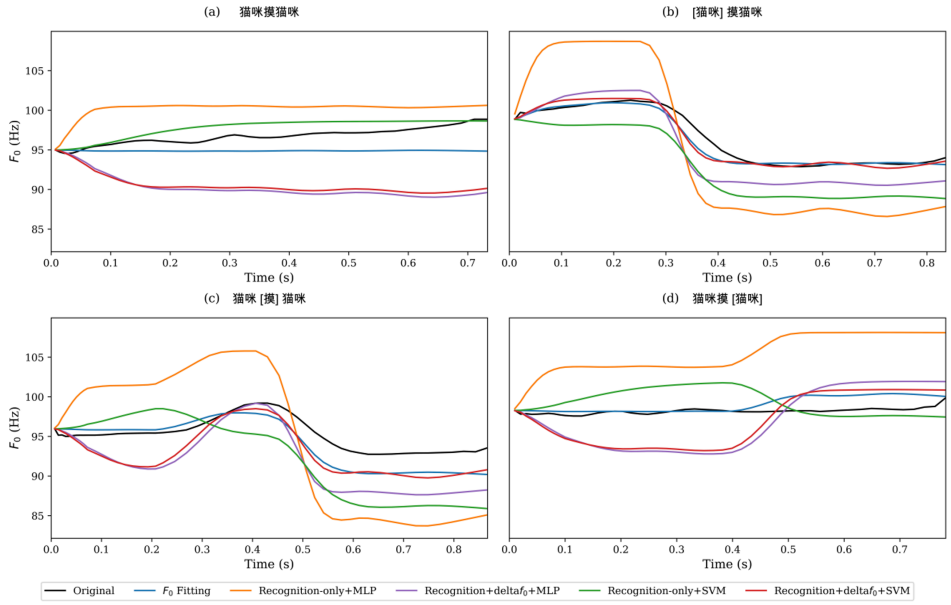



Figure 14: Original and synthetic $F_0$ contours of an all-High-tone sentence in four focus conditions)

## 5 Summary and conclusion

This study investigated how different learning strategies affect the modeling simulation of vocal learning of tone and focus in Mandarin. Through a comparison of $F_0$ fitting and recognition-guided learning with or without delta$F_0$ loss, we examined the impact of model architecture, training objective, and feature design on prosodic representation. The model performance was evaluated with three metrics: automatic recognition accuracy, human perception of tone and focus, and naturalness rating.

The results show that $F_0$ fitting had good performance as seen in previous studies (Xu & Prom-on 2014), achieving the best or second-best scores among all the models in tone recognition, focus identification, and perceived naturalness. Surprisingly, recognition-guided learning achieved largely equal perceptual identification of tone (with SVM, Figure 9) and focus (with MLP, Figure 11) as learning by $F_0$ fitting. Also, for recognition-guided models, the addition of delta$F_0$ as loss showed variable benefits across different tasks and model architectures. The MLP model benefitted from delta$F_0$ , while the SVM model showed trade-offs, with delta$F_0$ improving focus recognition but reducing tone accuracy. In Meng et al. (2023), delta$F_0$ was introduced to improve the $F_0$ fitting in terms of RMSE and correlation. The results of the current study show the limit of such improvement for human perception of tone and focus.

Moreover, the current study found that automatic evaluations match human judgments much more for tone perception than for focus perception (Figure 7/9 vs. Figure 8/11). Also interestingly, human judgment of naturalness was highly similar to that of tone perception accuracy across different sources of speech (Figure 9/13), but was unrelated to focus perception accuracy (Figure 11/13). This would raise questions about the usefulness of widely used mean opinion score (MOS) for assessing quality of synthetic speech, especially for the purpose of improving the prosody of speech synthesizers.

Taken together, the results of this study demonstrate, for the first time, the feasibility of simulating recognition-guided production learning of tone and focus at the same time. The target approximation parameters learned under recognition guidance achieved similar human perception accuracies as those learned by direct $F_0$ fitting. There is room for improvement, however, in terms of the amount and diversity of the training data, which can be explored in future research.

## References

't Hart, J., R. Collier & A. Cohen. 1990. *A perceptual study of intonation — an experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press.

Andruski, J. & J. Costello. 2004. Using polynomial equations to model pitch contour shape in lexical tones: An example from green mong. *Journal of the International Phonetic Association* 34. 125–140.

Bailly, G. & B. Holm. 2005. Sfc: A trainable prosodic model. *Speech Communication* 46. 348–364.

Barakat, Huda, Oytun Turk & Cenk Demiroglu. 2024. Deep learning-based expressive speech synthesis: A systematic review of approaches, challenges, and resources. *EURASIP Journal on Audio, Speech, and Music Processing* 2024(1). 11.

Boersma, Paul & David Weenink. 2001. Praat: doing phonetics by computer [Computer program]. *Glot International* 5(9-10). 341–347. DOI: 10 . 1126 / science . 6287572. http://www.praat.org/.

Botinis, A., M. Fourakis & B. Gawronska. 1999. Focus identification in English, greek and swedish. In *The 14th international congress of phonetic sciences*, 1557–1560. San Francisco.

Büring, Daniel. 2006. Focus projection and default prominence. In Valeria Molnar & Susanne Winkler (eds.), *The architecture of focus*.

Carlson, Katy, Michael Walsh Dickey, Lyn Frazier & Charles Clifton Jr. 2009. Information structure expectations in sentence comprehension. *Quarterly Journal of Experimental Psychology* 62(1). 114–139.

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

Chen, Yue, Yingming Gao & Yi Xu. 2022. Computational modelling of tone perception based on direct processing of f0 contours. *Brain Sciences* 12(3). 337.

Chen, Yue & Yi Xu. 2021. *Parallel recognition of mandarin tones and focus from continuous f0*.

Chen, Yue & Yi Xu. 2025. Sequential perception of tone and focus in parallel–a computational simulation. *Speech Communication* 168. 103173.

Cutler, A., D. Dahan & W. van Donselaar. 1997. Prosody in the comprehension of spoken language: A literature review. *Language and Speech* 40. 141–201.

Cybenko, George. 1989. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems* 2(4). 303–314.

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

Fujisaki, H. 1983. Dynamic characteristics of voice fundamental frequency in speech and singing. In P. F. MacNeilage (ed.), *The production of speech*, 39–55. New York: Springer-Verlag.

Fujisaki, H. 1992. Modeling the process of fundamental frequency contour generation. In Y. Tohkura, E. Vatikiotis-Bateson & Y. Sagisaka (eds.), *Speech perception, production and linguistic structure*, 313–326. Amsterdam: IOS Press.

Fujisaki, H. & S. Nagashima. 1969. *A model for the synthesis of pitch contours of connected speech*. Tech. rep. University of Tokyo.

Grabe, E., G. Kochanski & J. Coleman. 2007. Connecting intonation labels to mathematical descriptions of fundamental frequency. *Language and Speech* 50. 281–310.

Guenther, F.H., S.S. Ghosh & J.A. Tourville. 2006. Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language* 96. 280–301.

Hirst, Daniel. 2011. The analysis by synthesis of speech melody: From data to models. *Journal of Speech Sciences* 1. 55–83.

Hirst, Daniel J. 2004. Speech prosody: From acoustics to interpretation. *From Traditional Phonology to Modern Speech Processing (Festschrift for Professor Wu Zongji's 95th Birthday)*. 177–188.

Kreßel, Ulrich, Frank Lindner, Christian Wohler & Andreas Linz. 1999. Hypothesis verification based on classification at unequal error rates. In *9th international conference on artificial neural networks: Icann'99*, 874–879. IET.

Krug, P. K., P. Birkholz, B. Gerazov, D. R. van Niekerk, A. Xu & Y. Xu. 2023. Artificial vocal learning guided by phoneme recognition and visual information. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31. 1734–1744.

Kuhl, Patricia K. 2004. Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience* 5(11). 831–843.

Kuhl, Patricia K, Barbara T Conboy, Sharon Coffey-Corina, Denise Padden, Maritza Rivera-Gaxiola & Tobey Nelson. 2008. Phonetic learning as a pathway to language: New data and native language magnet theory expanded (nlm-e). *Philosophical Transactions of the Royal Society B: Biological Sciences* 363(1493). 979–1000.

Kuhl, Patricia K. 2000. A new view of language acquisition. *Proceedings of the National Academy of Sciences* 97(22). 11850–11857.

Kuhl, Patricia K. 2007. Is speech learning 'gated' by the social brain? *Developmental Science* 10(1). 110–120.

Ladd, D.R. 2008. *Intonational phonology*. Second. Cambridge: Cambridge University Press.

Lehiste, I. 1970. *Suprasegmentals*. MIT Press: Cambridge, MA.

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

Lieberman, Philip, Edmund S Crelin & Dennis H Klatt. 1972. Phonetic ability and related anatomy of the newborn and adult human, neanderthal man, and the chimpanzee. *American anthropologist* 74(3). 287–307.

Liu, F. 2009. *Intonation systems of mandarin and English: A functional approach*. The University of Chicago. (PhD dissertation).

Liu, Fang & Yi Xu. 2005. Parallel encoding of focus and interrogative meaning in mandarin intonation. *Phonetica* 62. 70–87.

Meng, Hanyi, Chien-tzu Chen, Yue Chen, Zirui Liu & Y. Xu. 2023. Mandarin tone production can be learned under perceptual guidance — a machine learning simulation. In *The 20th international congress of phonetic sciences*, 2324–2328. Prague, Czech.

Messum, Piers & Ian S Howard. 2015. Creating the cognitive form of phonological units: The speech sound correspondence problem in infancy could be solved by mirrored vocal interactions rather than by imitation. *Journal of Phonetics* 53. 125–140.

Ohman, S. & J. Lindqvist. 1965. *Analysis-by-synthesis of prosodic pitch contours.*

Palmer, H.E. 1922. *English intonation, with systematic exercises.* Cambridge: Heffer.

Panayotov, V., G. Chen, D. Povey & S. Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 ieee international conference on acoustics, speech and signal processing (icassp)*, 5206–5210.

Pierrehumbert, J. 1981. Synthesizing intonation. *Journal of the Acoustical Society of America* 70. 985–995.

Pike, Kenneth L. 1945. *The intonation of American English.* Ann Arbor: University of Michigan Press.

Prom-on, Santitham & Yi Xu. 2012. Pentatrainer2: A hypothesis-driven prosody modeling tool. In *Exling 2012*, 93–100. Athens, Greece.

Prom-on, Santitham & Yi Xu. 2013. Modeling speech melody as communicative functions with pentatrainer2. In *Tools and resources for the analysis of speech prosody (trasp 2013)*, 82–85. Aix-en-Provence, France.

Prom-on, Santitham, Yi Xu & Bundit Thipakorn. 2009. Modeling tone and intonation in Mandarin and English as a process of target approximation. *The Journal of the Acoustical Society of America* 125(1). 405–24. DOI: 10.1121/1.3037222.

Purcell, David W & Kevin G Munhall. 2006. Compensation following real-time manipulation of formants in isolated vowels. *The Journal of the Acoustical Society of America* 119(4). 2288–2297.

Reddy, V Ramu & K Sreenivasa Rao. 2016. Prosody modeling for syllable based text-to-speech synthesis using feedforward neural networks. *Neurocomputing* 171. 1323–1334.

Rump, H.H. & R. Collier. 1996. Focus conditions and the prominence of pitch-accented syllables. *Language and Speech* 39. 1–17.

Taylor, P. 2000. Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America* 107. 1697–1714.

Taylor, Paul. 2009. *Text-to-speech synthesis*. Cambridge: Cambridge University Press.

Tsao, Feng-Ming, Huei-Mei Liu & Patricia K Kuhl. 2004. Speech perception in infancy predicts language development in the second year of life: A longitudinal study. *Child development* 75(4). 1067–1084.

Van Niekerk, Daniel R., Anqi Xu, Branislav Gerazov, Paul K. Krug, Peter Birkholz, Lorna Halliday, Santitham Prom-on & Yi Xu. 2023. Simulating vocal learning of spoken language: Beyond imitation. *Speech Communication* 147. 51–62.

Vihman, Marilyn May & Bénédicte de Boysson-Bardies. 1994. The nature and origins of ambient language influence on infant vocal production and early words. *Phonetica* 51(1-3). 159–169.

Wagner, Michael & Duane G. Watson. 2010. Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes* 25(7). 905–945.

Werker, Janet F. & Richard C. Tees. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development* 7(1). 49–63.

Xu, Anqi, Daniel R. van Niekerk, Branislav Gerazov, Paul Konstantin Krug, Peter Birkholz, Santitham Prom-on, Lorna F. Halliday & Yi Xu. 2024. Artificial vocal learning guided by speech recognition: What it may tell us about how children learn to speak. *Journal of Phonetics* 105. 101338.

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

Xu, Y. 1999. Effects of tone and focus on the formation and alignment of f0 contours. *Journal of Phonetics* 27. 55–105.

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

Xu, Y., Szu-wei Chen & Bei Wang. 2012. Prosodic focus with and without post-focus compression (pfc): A typological divide within the same language family? *The Linguistic Review* 29. 131–147.

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

Xu, Yi. 2019. Prosody, tone and intonation. In W. F. Katz & P. F. Assmann (eds.), *The routledge handbook of phonetics*, 314–356. Philadelphia: Routledge.

Xu, Yi & Santitham Prom-on. 2014. Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. *Speech Communication* 57. 181–208.

Xu, Yi, Anqi Xu, Daniel R. van Niekerk, Branislav Gerazov, Peter Birkholz, Paul Konstantin Krug, Santitham Prom-on & Lorna F. Halliday. 2022. Evoc-learn — high quality simulation of early vocal learning. In *Interspeech 2022*.

Yan, Mengzhu & Sasha Calhoun. 2020. Rejecting false alternatives in Chinese and English: The interaction of prosody, clefting, and default focus position. *Laboratory Phonology* 11(1).

Yuan, Jiahong. 2011. Perception of intonation in mandarin Chinese. *The Journal of the Acoustical Society of America* 130(6). 4063–4069.