INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

An Investigation of the Target Approximation Model for Tone Modeling and Recognition in Continuous Mandarin Speech

Yingming Gao[1], *Xinyu Zhang*[1], *Yi Xu*[2], *Jinsong Zhang*[3], *Peter Birkholz*[1]

[1]Institute of Acoustics and Speech Communication, TU Dresden, Germany
[2]Department of Speech, Hearing and Phonetic Sciences, University College London, UK
[3]College of Information Sciences, Beijing Language and Culture University, China

yingming.gao@mailbox.tu-dresden.de, duanfengzxy@gmail.com, yi.xu@ucl.ac.uk, jinsong.zhang@blcu.edu.cn, peter.birkholz@tu-dresden.de

Abstract

The complex f_0 variations in continuous speech make it rather difficult to perform automatic recognition of tones in a language like Mandarin Chinese. In this study, we tested the use of target approximation model (TAM) for continuous tone recognition on two datasets. TAM simulates f_0 production from the articulatory point of view and so allow to discover the underlying pitch targets from the surface f_0 contour. The f_0 contour of each tone represented by 30 equidistant points in the first dataset was simulated by the TAM model. Using a support vector machine (SVM) to classify tones showed that, compared to the representation by 30 f_0 values, the estimated three-dimensional TAM parameters had a comparable performance in characterizing tone patterns. The TAM model was further tested on the second dataset containing more complex tonal variations. With equal or a fewer number of features, the TAM parameters provided better performance than the coefficients of the cosine transform and a slightly worse performance than the statistical f_0 parameters for tone recognition. Furthermore, we investigated bidirectional LSTM neural network for modelling the sequential tonal variations, which proved to be more powerful than the SVM classifier. The BLSTM system incorporating TAM and statistical f_0 parameters achieved the best accuracy of 87.56%.

Index Terms: continuous Mandarin speech, tone modeling and recognition, target approximation model, LSTM neural network

1. Introduction

Mandarin Chinese (Standard Chinese) is a well-known syllable-based tone language. Each syllable is associated with one of five pitch tones, including four lexical tones (referred to as Tones 1-4) and a neutral tone (referred to as Tone 5). The 412 base syllables stratified with five tones constitute about 1282 phonetically differentiated syllables in Mandarin. Pitch tones play crucial phonemic roles so that the same syllable with different tones has different lexical meanings. Therefore, automatic Mandarin tone recognition is a fundamental and nontrivial research topic, which benefits discriminating homophone words in automatic speech recognition (ASR) systems, labeling prosodic information of databases for data-driven text-to-speech (TTS) systems, and detecting tone pronunciation errors for computer aided language learning (CALL) systems.

The four lexical tones are phonologically characterized by the f_0 patterns, namely, Tone 1 (high-level), Tone 2 (mid-rising), Tone 3 (falling-rising), and Tone 4 (high-falling). When pronounced in isolation, their pitch contours seem well defined and quite stable (demonstrated as dotted lines in Figure 1). The tone recognition of isolated syllables is relatively easy, achieving an accuracy of above 99% [1] for four lexical tones.

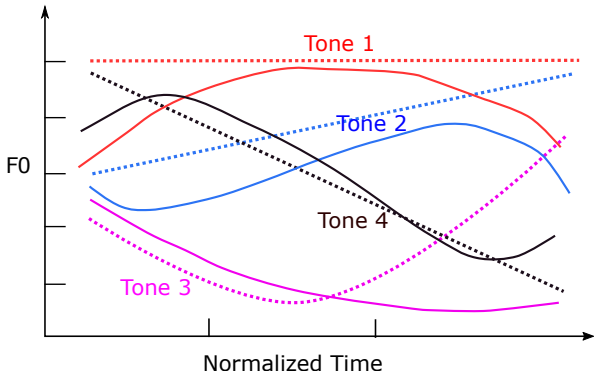


Figure 1: *Illustration of standard f_0 patterns of the four lexical tones (dotted ones) and their possible variations (solid ones) when affected by specific neighboring tones in continuous speech (adapted from Chao 1968 [2] and Zhang et al. 2005 [3]). Note: they are not quantitative but qualitative and descriptive. These tone variants demonstrate some frequently observed patterns with possible transitions affected by adjacent syllables.*

However, these standard tonal patterns seldom occur in continuous speech due to complex physiological, phonetic, and linguistic constraints etc. First, the tonal co-articulation influences the pitch contour. Being affected by the tones of neighbouring syllables, the surface F0 contours are usually undershot. As the solid contours show in Figure 1, the tone patterns of continuous speech deviate a lot from their standard ones. Moreover, tones are embedded in the intonation structure of the whole utterance. For example, sentence type, speaking style, and topic shift can result in complex f_0 variations, because the local prosodic units (tones) are modulated so as to accommodate to the global prosodic structure. Besides, Tone 5, also known as the neutral tone, has greater contextual variability than the lexical tones because its f_0 varies greatly with the preceding tone. All these tonal variations increase the difficulties of tone modeling and recognition in continuous Mandarin speech.

Regarding automatic tone recognition, researchers attempted many approaches in terms of different kinds of features and classifiers. For features, traditional methods usually use prosodic features e.g. f_0 and energy as well as their derived features. In addition to the prosodic features, more recent methods also employed spectral features due to the potential correlation between source features and spectral features, e.g. mel-frequency cepstral coefficients (MFCCs) used by Ryant et al.

http://dx.doi.org/10.21437/Interspeech.2020-2823

[4, 5]. Chao et al. [6] and Lin et al. [7] also incorporated articulatory features into tone recognition. For classifiers, researchers have tried MLP, SVM, DNN, etc. [6, 7, 8]. To employ the contextual and temporal information, researchers usually used HMM [7] or LSTM models [8].

However, those data-driven methods did not take the f_0 generation mechanism into consideration. Zhang et al. [3, 9] proposed to divide the f_0 contour of a syllable into tone nucleus and adjacent articulatory transitions, then used the tone nucleus as the underlying tone target for recognition. Wang et al. [10] used phrasing coefficients of the Fujisaki model [11] as features for tone recognition. Prom-on et al. [12] developed the quantitative implementation of target approximation model (TAM) for generating the f_0 contour of English and Mandarin speech, which was later successfully tested for Thai and German f_0 modeling [13, 14]. The TAM generates surface f_0 contours from invariant underlying pitch targets based on articulatory dynamics. The relative invariance of underlying pitch targets may help to distinguish tones from complex surface f_0 contours. In this study, we propose to apply qTA to model continuous Mandarin f_0 contours, and then use the derived parameters as the features for tone recognition.

2. Method

The goal of this study is to explore the benefit of the TAM for tone modeling from raw pitch contours for continuous Mandarin speech and its application to tone recognition. Although elaborated features (e.g. spectral features) or classifiers may improve the performance, the experiments are focused on how effective the TAM is to abstract the raw f_0 contour and how distinct the derived parameters are for tone recognition. Any recognition system built with other features and classifiers can easily incorporate it.

2.1. Target approximation model

The target approximation model (TAM), as illustrated in Figure 2, attempts to articulatorily simulate the underlying mechanisms of f_0 realization. It assumes one pitch target for each syllable of an utterance. Within the interval of a syllable, the target $x(t)$ is defined as the linear function,

$$x(t) = mt + b \tag{1}$$

The $f_0(t)$ is modeled as the response of an N-order critically damped linear system driven by a pitch target, as shown in the following equation,

$$f_0(t) = (mt + b) + (c_0 + c_1 t + ... + c_{N-1} t^{N-1}) e^{-t/\tau} \tag{2}$$

where m (in st/s) and b (in st) denote the slope and offset of the underlying pitch target, respectively, and the time constant τ (in s) represents the strength of the target approximation movement. In general, positive and negative values of m indicate rising and falling targets, respectively, while positive and negative values of b indicate raising and lowering of pitch targets relative to the speaker average f_0 level, respectively. The values of these three parameters are expected to reflect the tonal patterns. Applying TAM to Mandarin tones, Prom-on et al. [12] found, for example, that the Mandarin rising (R) and falling (F) tones have positive and negative m values, respectively, and that neutral tone have relatively greater τ values than those of other tones. Moreover, Xu et al. [13] analyzed 1280 eight-syllable Mandarin utterances and discovered the many-to-one mappings between surface f_0 and underlying representations. i.e., the invariant targets of variable surface f_0.

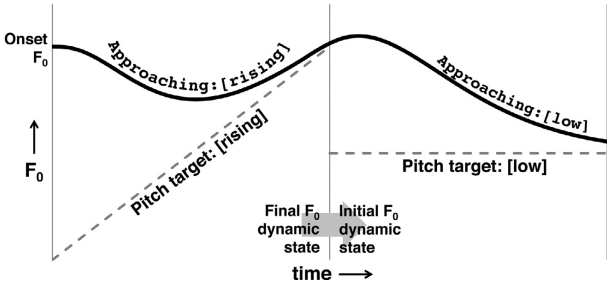


Figure 2: *An illustration of the target approximation process. The surface f_0 contour (indicated by the thick solid line) asymptotically approaches two underlying pitch targets (represented by the dashed lines). The middle vertical gray line represents the syllable boundary through which the final f_0 dynamic state is transferred from one syllable to the next. Source: Figure 2 in Xu et al. 2014 [13].*

2.2. Estimation of pitch targets

To estimate the underlying targets of tones, we adopted TargetOptimizer [14], one of the quantitative implementations of TAM. TargetOptimizer jointly estimated the pitch targets $\mathbf{p_s} = (m_s, b_s, \tau_s)^{\mathrm{T}}$, referred to as qTA parameters, for all syllables of the utterances by minimizing the following objective function,

$$\begin{aligned} g\left(\mathbf{p}_1 \dots \mathbf{p}_S\right) = & \left\| f_0(k\Delta t) - \hat{f}_0\left(k\Delta t, \mathbf{p}_1 \dots \mathbf{p}_S\right) \right\|_2^2 \\ & + \lambda \sum_{s=1}^{S} \left(\mathbf{p}_s - \overline{\mathbf{p}}\right)^T W \left(\mathbf{p}_s - \overline{\mathbf{p}}\right) \end{aligned} \tag{3}$$

where S is the number of syllables. The first term is the squared Euclidean distance between the original f_0 samples and the reproduced $\hat{f}_0$ samples by the TAM, which are sampled every $\Delta t = 10$ ms and indexed by k. The second part in Equation 3 is the regularization term, which penalized the deviation of qTA parameters from their preferred values $\overline{\mathbf{p}} = (\overline{m}, \overline{b}, \overline{\tau})^{\mathrm{T}}$. $W = diag(w_m, w_b, w_\tau)$ is the weight, the values of which were optimized in the experiments. λ is the overall regularization factor (see Birkholz et al. [14] for details). The only required annotation for using TargetOptimizer is the syllable boundaries, which are assumed available from a process of phoneme recognition or phoneme boundary detection.

2.3. Tone classifier

We employed two classifiers to recognize Mandarin tones: support-vector machine (SVM) and bidirectional Recurrent Neural Network with Long Short-Term Memory units (BLSTM). The SVM has less risk of over-fitting and no strict requirement for the size of dataset. It is easy to train and build as a baseline system. The second classifier we used is the BLSTM, which has demonstrated its capability of modeling temporal sequences in tasks of handwriting recognition, language translation, speech recognition and so on. The reason for adopting BLSTM is to deal with the bidirectional contextual effects (*carry-over* and *anticipatory* effects from neighboring tones) and even long distant prosodic effects.

3. Evaluation

3.1. Speech corpora

3.1.1. Dataset-1: Controlled speech

The controlled speech was produced by four female and four male Chinese speakers [15]. Each of 16 disyllabic nonsense sequences with balanced tonal structures was spoken by each speaker in four different carrier sentences. The pitch contour taken from either the first or second syllable of a disyllabic target word was sampled to 30 equidistant (hence time-normalized) discrete points. This dataset concerned only four lexical tones. To simplify the recognition task, the Tone3-Tone3 cases were excluded because the first Tone 3 has become nearly identical to Tone 2 due to a phonological tone sandhi rule. There were a total of 1408, 1408, 1232 and 1408 tokens for Tone 1, Tone 2, Tone 3 and Tone 4, respectively.

3.1.2. Dataset-2: Uncontrolled speech

The uncontrolled speech used in this study is the native part of the BLCU inter-Chinese speech corpus [16], in which each of the 12 Chinese native speakers (six females and six males) produced 301 sentences. The corpus contains phoneme and syllable labels together with orthographic transcriptions. The average length per utterance is 7.7 syllables. The average speaking rate over all utterances is 5.22 syllables per second. Compared to the Dataset-1, the speech of this dataset contained more tonal variations, thus increasing the difficulty of tone recognition. Besides, Dataset-2 included the Tone3-Tone3 cases, the first of which was modified to Tone 2 according to the tone sandhi rule during subsequent tone modeling and recognition. The neutral tones were also kept.

3.2. Feature Extraction

The utterances in Dataset-2 were split into shorter ones when they contained silence. The f_0 contours were extracted from the audio files using the auto-correlation method of Praat [17] with a time step of 10 ms. The preferred values of TargetOptimizer were set to $\overline{m} = 0$, $\overline{\tau} = 20$ ms and $\overline{b}$ to the average pitch of the utterances. We selected 1000 utterances from Dataset-2 for optimizing the hyperparameters of TargetOptimizer with the grid search method. With the least root mean square error (RMSE) between the original pitch contour and reproduced one as the metric and the consideration of reasonable distributions of estimated qTA parameters, we obtained the best configuration of TargetOptimizer: 0.01 $\mathrm{s}^2/\mathrm{st}^2$ for w_m, 0.01 st^{-2} for w_b, and 5 s^{-2} for w_τ. Subsequently, the qTA parameters of all utterances for both corpora were extracted. The utterances in Dataset-2 ($\sim$ 5.4%) were excluded when the TargetOptimizer did not converge during optimization or the RMSE was greater than 1.5 semitones. The final Dataset-2 contained 5027 Tone 1s, 5749 Tone 2s, 5620 Tone 3s, 8064 Tone 4s, and 3479 Tone 5s.

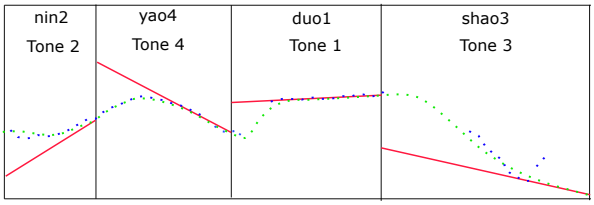
The averaged TAM parameters extracted for the four lexical tones in Dataset-1 are listed in Table 1. The slope characterized the asymptotic direction of pitch contour for each tone, which is consistent with the finding by Prom-on et al. [12]. An interesting phenomenon happened to the offset value. The high pitch offset of Tone 4 is consistent with the finding of Xu [15] where this tone is the highest among the four tones in Mandarin. To examine whether four tones are contrastive with each other from statistical point of view, we conducted the multivariate analysis of variance (MANOVA) with three qTA parameters as the dependent variables and tone types as the independent variables. The results showed significant differences of qTA parameters for each tone pair, suggesting the possibility of using them for tone recognition.

Table 1: *Averaged TAM parameters for four lexical tones in Dataset-1 (standard deviations in parenthesis).*

Tone type	Slope	Offset	Time constant
Tone 1	1.40 (4.97)	91.90 (7.32)	19.64 (1.46)
Tone 2	15.12 (9.30)	85.71 (7.07)	20.09 (1.93)
Tone 3	−14.08 (17.21)	87.44 (6.87)	21.60 (3.51)
Tone 4	−20.30 (11.66)	95.45 (7.32)	19.57 (1.78)

The same procedure was also applied to Dataset-2. An example of modeling the f_0 contour of a continuous Chinese utterance with TargetOptimizer is shown in Figure 3. The blue dotted loci are the raw f_0 samples extracted with Praat. The green dotted contour is reproduced via TargetOptimizer with the estimated pitch targets represented by red solid lines. As we can see from Figure 3, although the surface f_0 contour of each syllable varies a lot in the beginning part, the end of it gradually approximates its underlying pitch target. Therefore, expressing the variable with the parameters of underlying pitch targets may help distinguish tones from each other.

Moreover, for Dataset-2, we compared the TAM modeling with two other ways of representing pitch information. The first was to use statistical f_0 parameters, including the onset, offset, maximal, minimal, mean, median, and standard deviation values. To reflect the transition information from/to neighbouring syllables, we incorporated the offset values of its preceding syllable and the onset value of its succeeding syllable. These values were subtracted by the sentence mean f_0 value so as to alleviate the influence of gender pitch difference. Likewise, the offset of TAM was changed to a relative value in this way. The second was to first conduct the Discrete Cosine Transform (DCT) on the raw f_0 values of each syllable, and then select the first three coefficients to characterize the pitch contour of this syllable. As the neutral tone has no specific pattern and varies greatly with contextual tone types, the pitch information can not be fully represented by the above three kinds of methods. According to the report by Chen et al. [18], the neutral tone is significantly shorter than lexical tones. Therefore, we included the duration as a complementary feature to the above three sets of parameters.



parameter	target[1]	target[2]	target[3]	target[4]
slope [st/s]	60.000	-50.218	5.089	-21.515
offset [st]	88.204	103.760	98.257	92.174
tau [ms]	14.052	20.127	8.460	24.952
duration [s]	0.132	0.190	0.212	0.296

Figure 3: *Illustration of modeling f_0 contour with TargetOptimizer (With the utterance "nin2 yao4 duo1 shao3?" spoken by a female speaker as an example).*

3.3. Tone recognition

Using SVM classifiers, we examined the capability of qTA parameters for tone recognition on Dataset-1, which was divided into a training subset and a testing subset with a ratio of 3:1. Using the LIBSVM library [19], we trained an RBF kernel based SVM classifiers with grid search for hyperparameters (the cross-product of exponentially growing sequences of $C \in \{2^{-5}, 2^{-3}, ..., 2^{15}\}$ and $\gamma \in \{2^{-15}, 2^{-13}, ..., 2^{3}\}$). Here, we developed three systems contrasted by the input features used. The recognition results for the four lexical tones are listed in Table 2. The first used the sampled 30 dimensional raw f_0 points, which was taken as the targeted performance system. The qTA parameters (m, b, and τ) only represent the pitch targets of tones. The transition status from the preceding syllable is also required to generate the pitch contours of the current syllable according to the TAM model, and then expected to be helpful to tone recognition. Therefore, the third system further incorporated the onset value of each syllable upon the second system. The results showed that the pitch targets conveyed most information for distinguishing tones. With the additive onset value, the parameters of pitch targets obtained similar performance, compared to the way of representing pitch contour by 30 f_0 samples.

Table 2: *SVM classifier based tone recognition accuracy on Dataset-1.*

Features used	Accuracy
sampled 30-dim f_0 points	97.42%
3-dim qTA parameters	90.67%
3-dim qTA parameters plus onset value	97.07%

Using BLSTM and SVM classifiers, we compared the capability of three kinds of features. The BLSTM based tone classifiers were implemented with PyTorch [20]. The models were trained for 50 epochs in which weights and bias were updated by the Adam optimization algorithm [21]. For each miniBatch, we sorted all utterances by their lengths (i.e., the number of syllables) in a descending order. Then, we padded zeros to short utterances so that all utterances had the same length. Hyperparameters for BLSTM neural networks were optimized with the grid search strategy from all combinations of other hyperparameters values (the number of layers: {1, 2, 3}, the number of units per layer: {300, 400, 500, 600}, the batch size: {16, 32, 64, 128}, the initial learning rate: {0.001, 0.002, 0.005, 0.01, 0.015}). To use more samples during hyperparameters optimization of BLSTM classifiers, we adopted the repeated random sub-sampling validation where we randomly split the whole dataset into training and validation data (80% vs. 20%), which were used for fitting and assessing the model, respectively. We repeated this procedure five times for each set of hyperparameters and averaged the accuracies of all repetitions. The results of recognition accuracy on Dataset-2 are listed in Table 3. Syllable duration was used by all systems, thus not explicitly differentiating the systems.

As shown in Table 3, the statistical f_0 features had the general better performance, especially among the SVM based classifiers. This may be explained by the fact that they explicitly included information from neighbouring syllables, i.e. the offset of the preceding tones and the onset of the succeeding tones. The BLSTM based systems outperformed all SVM based systems, indicating its advantages of modeling contextual tonal variation. The long-term memory of BLSTM may deal with the variation due to the intonation while its short-term memory may tackle the variation due to influences of neighbouring tones. Specifically, the *bidirectional* attribute of BLSTM may account for the *carry-over* and *anticipatory* effects. To examine whether the information that the qTA parameters represent is also included in the statistical f_0 parameters, we developed another system by incorporating both of them. The further improvement indicates that the qTA parameters do contain additive information. For example, the τ of qTA determines how quickly the transition from one syllable to the next completes, which is not reflected by the statistical f_0 parameters. The confusion matrix for the best system is shown in Table 4.

Table 3: *Tone recognition accuracy on Dataset-2.*

Classifier and features	No. of features	Accuracy
SVM + DCT coefficients	4	54.53%
SVM + qTA parameters	4	55.76%
SVM + f_0 statistics	10	64.08%
BLSTM + DCT coefficients	4	78.58%
BLSTM + qTA parameters	4	84.32%
BLSTM + f_0 statistics	10	85.27%
BLSTM + f_0 + qTA parameters	13	87.56%

Table 4: *Confusion matrix for tone recognition on Dataset-2 using BLSTM classifier and 13-dim combined parameters.*

	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5
Tone 1	**85.3%**	3.93%	2.28%	6.48%	2.03%
Tone 2	2.77%	**86.67%**	4.2%	3.84%	2.53%
Tone 3	1.76%	3.2%	**87.03%**	5.34%	2.66%
Tone 4	3.81%	3.53%	3.18%	**88.1%**	2.38%
Tone 5	2.82%	4.33%	4.48%	5.47%	**82.90%**

4. Conclusions and future work

In the present study, we evaluated the the effectiveness of applying TAM to Chinese tone modeling and recognition on two datasets. The experiments showed that, with only three dimensions, the qTA parameters of pitch target can convey distinctive information for tone recognition. Compared to the SVM model, the BLSTM significantly increased the tone recognition accuracy, indicating its effectiveness on modeling temporal serial data. The proposed method was validated on two speech datasets containing human-labeled syllable segmentation. Although the qTA model proved to be non-sensitive to boundaries [14], it is necessary to further verify the approach using automatically derived segmentation, such as from phoneme boundary detection. The hyperparameters of TargetOptimizer were optimized with the goal of least RSME, which may drive the reproduced pitch contours to *overfit* the original ones. The underlying pitch targets estimated by such hyperparameters may not be the optimal for tone recognition. Future work will also be focused on accurately estimating the underlying pitch targets.

5. Acknowledgements

This study was partially supported by China Scholarship Council and the Science Foundation of Beijing Language and Culture University (the Fundamental Research Funds for the Central Universities, 20YJ040002).

6. References

[1] Q. Gao, S. Sun, and Y. Yang, "Tonenet: A cnn model of tone classification of mandarin chinese." in *Proc. Interspeech 2019*, 2019, pp. 3367–3371.

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

[3] J. Zhang, S. Nakamura, and K. Hirose, "Tone nucleus-based multi-level robust acoustic tonal modeling of sentential f0 variations for chinese continuous speech tone recognition," *Speech Communication*, vol. 46, no. 3-4, pp. 440–454, 2005.

[4] N. Ryant, J. Yuan, and M. Liberman, "Mandarin tone classification without pitch tracking," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4868–4872.

[5] N. Ryant, M. Slaney, M. Liberman, E. Shriberg, and J. Yuan, "Highly accurate mandarin tone classification in the absence of pitch information," in *Proceedings of Speech Prosody*, vol. 7, 2014.

[6] H. Chao, Z. Yang, and W. Liu, "Improved tone modeling by exploiting articulatory features for mandarin speech recognition," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4741–4744.

[7] J. Lin, W. Li, Y. Gao, Y. Xie, N. F. Chen, S. M. Siniscalchi, J. Zhang, and C.-H. Lee, "Improving mandarin tone recognition based on dnn by combining acoustic and articulatory features using extended recognition networks," *Journal of Signal Processing Systems*, vol. 90, no. 7, pp. 1077–1087, 2018.

[8] L. Yang, Y. Xie, and J. Zhang, "Improving mandarin tone recognition using convolutional bidirectional long short-term memory with attention." in *Proc. Interspeech 2018*, 2018, pp. 352–356.

[9] J. Zhang and K. Hirose, "Tone nucleus modeling for chinese lexical tone recognition," *Speech Communication*, vol. 42, no. 3-4, pp. 447–466, 2004.

[10] C. Wang, H. Fujisaki, and K. Hirose, "Chinese four tone recognition based on the model for process of generating f0 contours of sentences," in *First International Conference on Spoken Language Processing*, 1990, pp. 221–224.

[11] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of japanese," *Journal of the Acoustical Society of Japan (E)*, vol. 5, no. 4, pp. 233–242, 1984.

[12] S. Prom-On, Y. Xu, and B. Thipakorn, "Modeling tone and intonation in mandarin and english as a process of target approximation," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 405–424, 2009.

[13] Y. Xu and S. Prom-On, "Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning," *Speech Communication*, vol. 57, pp. 181–208, 2014.

[14] P. Birkholz, P. Schmager, and Y. Xu, "Estimation of pitch targets from speech signals by joint regularized optimization," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2075–2079.

[15] Y. Xu, "Contextual tonal variations in mandarin," *Journal of Phonetics*, vol. 25, no. 1, pp. 61–83, 1997.

[16] W. Cao, D. Wang, J. Zhang, and Z. Xiong, "Developing a chinese l2 speech database of japanese learners with narrow-phonetic labels for computer assisted pronunciation training," in *Proc. Interspeech 2010*, 2010, pp. 1922–1925.

[17] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," *http://www.praat.org*, 2020.

[18] Y. Chen and Y. Xu, "Production of weak elements in speech–evidence from f patterns of neutral tone in standard chinese," *Phonetica*, vol. 63, no. 1, pp. 47–75, 2006.

[19] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[20] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.