A Preliminary Study on Analysing Mandarin Tone Values of Romance L2 Mandarin Learners

Wu-Hao Li*, Te-hsin Liu† and Chen-Yu Chiang‡
* College of Electrical Engineering and Computer Science, National Taipei University, Taiwan
E-mail: hank12451@gmail.com Tel/Fax: +886-2671-0893
† Graduate Program of Teaching Chinese as the Second Language, National Taiwan University
E-mail: tehsinliu@ntu.edu.tw Tel/Fax: +886-3366-1549
‡ Dept. of Communication Engineering, National Taipei University, New Taipei City, Taiwan
E-mail: cychiang@mail.ntpu.edu.tw Tel/Fax: +886-2671-0893

***Abstract*—The 5-level tone value labeling system helps characterize pitch contours of syllables for L2 Mandarin learners and teachers to facilitate tone acquisition and tone error analysis, respectively. In the past, linguistic experts did the tone value labeling and initially used it to represent or analyze the pitch contours of monosyllables. Some L2 Mandarin teachers also used the tone value labeling system to analyze pitch contours of continuous speech and illustrate pitch realizations for L2 Mandarin learners. However, the pitch contours of the labeled tone values of continuous speech were rarely objectively measured or analyzed. This paper presents a preliminary analysis of tone values of a continuous read speech corpus uttered by six Romance L2 Mandarin learners. The pitch contours of the tone value classes were explored, and some experiments related to the automatic recognition of tone and tone value classes were reported. We expect the results reported by this study to shed light on the development of an automatic tone value classification mechanism, which will be used in computer-assisted pronunciation training for L2 learners in the future.**

***Index Terms*—Mandarin, tone recognition, tone value, computer-assisted pronunciation training**

I. Introduction

L2 Mandarin learners would utter syllables with incorrect tones. These incorrect tones would be affected by learners' mother tongues and sound differently from native Mandarin speakers. The most intuitive method to access the tone uttered by L2 Mandarin learners is to utilize the results of a machine tone recognizer. The tone recognition results can be converted to some metrics as measures for tone assessments or error detection [1]–[3]. However, the tone recognition results, e.g., the posterior probabilities of tones, do not easily illustrate how the pitch contours of speech uttered by L2 Mandarin learners differ from native Mandarin speakers' tones. For example, if the tone recognizer produces the score for a target Mandarin syllable with scores of 0.1, 0.2, 0.3, 0.3, and 0.1 for tones 1 to 5, this score information is vague for an L2 Mandarin learner to change his/her produced pitch contour into correct one.

For linguistic background researchers or L2 Mandarin teachers, using the 5-level tone values to characterize pitch contours of speech uttered by L2 Mandarin learners is more intuitive or more comprehensive than using tone classes. The 5-level tone values were first introduced by Chao [4], [5] to initially classify the four lexical tones in Mandarin as typical tone value sequences of 55, 35, 214, and 51, for T1, T2, T3, and T4, respectively. In most of the previous studies or teaching materials [6], [7], the tone value transcriptions were based on human ear annotations, supplemented by pitch contours. For example, L2 Mandarin learners may utter a T1 syllable with not a high-enough 44 tone value in the sequence, which has a lower level than the standard T1 with a tone value of 55.

Since the labeling involves human perception in audio and vision, the tone values labeled may suffer from inconsistencies between labelers. Though each utterance can be labeled by many human labelers with consistency checked by discussion between the labelers, the discussion process would be time-consuming and still involve human subjective biases. Besides labeling tone value by perception, Shi [8] proposed mapping from logF0 to Chao's 5-level tone values with maximum and minimum logF0 of speakers. Edmonson [9] proposed to first map from measured logF0 to semitone and then to Chao's 5-level tone values. The successive studies [10]–[12] generally follow the formula proposed by Shi and Edmonson and modify the formula regarding measuring methods to remove or normalize inter-speaker pitch differences.

The formulas proposed by Shi [8] and Edmonson et al. [9] were mainly used to represent tone values of lexical tones in monosyllables. Therefore, the above-mentioned 5-level tone representation and their variations for normalization would no longer be suitable to label the tone values of syllables in a continuous speech due to prosodic variation made by coarticulation and prosodic phrasing though many studies still use this 5-level tone value system to annotate syllable tones of continuous Thai [13], [14] speech.

This paper presents a preliminary analysis of tone values of continuous read speech uttered by six romance L2 Mandarin learners. The tone values of syllables are labeled by linguistic experts with the 5-scale values. To objectively visualize the tone values labeled, we first stylize the pitch contours of syllables by polynomial basis functions, which describe the mean, slope, acceleration, and curvature of syllable logF0 contour. We, then analyze the typical patterns of each tone value with the average observed logF0 contours represented by the coefficients of the polynomial functions. Last, we conducted automatic machine tone and tone value classification experiments to obtain baselines for tone and tone value classification tasks.

Contributions made by this paper are summarized in two folds. First, typical logF0 patterns for tone values in continuous speech uttered by L2 Mandarin learners are objectively explored. Most previous literature related to tone values considered primarily illustrating pitch contours of tones, especially for pronouncing monosyllables. Second, this paper reports the improvement of the tone recognition accuracy with respect to the result reported in 2023 [15]. Based on this improvement, this paper constructs the baseline for the performance of the tone value class recognition. So far, to our knowledge, this is the first paper to report the performance of tone value recognition.

This paper is organized as follows: Section II introduces the experimental databases used in this study. Section III analyzes the pitch characteristics of tone classes and tone value classes. Section IV conducts tone class and tone value class recognition experiments. The last section reports some conclusions and future works.

II. Experimental Databases

A. Romance L2 Mandarin Speech Corpus

1) Recording and Labeling: The Romance L2 Mandarin speech corpus is used as the material for this preliminary study. The dataset contains six (3 males and 3 females) French learners of Mandarin who participated in creating audio recordings. The participants were second-year (intermediate level) students majoring in Chinese at Institut National des Langues et Civilisations Orientales in Paris. None of the students had previous experience with other tonal languages. The participants were asked to read two short texts (completable within 5.5 min on average) at a natural speed. The first was a brief introduction to the Lunar New Year; the second was an anecdote from a high school student written in the first person.

There are a total of 4,381 syllables in this corpus, and each syllable is labeled with the 5-level tone values by three native Mandarin-speaking linguistic experts in linguistics. After the discussions by the three linguistic experts, the inconsistent or conflict tone value labeling was reduced as much as possible. There are 41 types of tone values labeled by the three experts. For labeling tone classes, each syllable of the corpus was first labeled with a ***lexical tone*** by the linguistic processor of the Speech Labeling and Modeling Toolkit (SLMTK) Version 1.0 [16]. Then, three other native Mandarin speakers labeled the corpus's syllables with perceived tones using the following procedure: the lexical tone labels for syllables were first provided to the three labelers as references, and the labelers relabeled the syllables with the tone classes that they heard if their perceived tone classes were different from the lexical tones.

2) Consistencies of Tone and Tone Value Labelings: Table I shows the percentages for the numbers that denote how many different labels a syllable received. It is found that the consistency in tone class labeling is higher than the consistency in tone value class labeling, maybe because the number of classes for the tone value labeling, i.e., 41, is larger than the number of tone classes, i.e., 5. Table II displays the correlation matrix for the lexical-perceived tone pairs. It is found that syllables with lexical tone 3 would be easily perceived as tone 2 due to tone three Sandhi. Except for tone three Sandhi, the percentages not on the diagonal may indicate the erroneous tones produced by the six L2 Mandarin learners.

TABLE I: percentages for the numbers 1 to 3 that denote how many different labels a syllable received. For example, 1 means all three labelers assigned the same tone tag to a syllable.

	1	2	3
tone classes	80.8%	12.0%	7.2%
tone value classes	60.0%	30.4%	9.6%

TABLE II: Correlation matrix for lexical and perceived tones: The data annotated by three labelers were combined, resulting in a dataset three times larger than the original.

		perceived tones (%)					
		1	2	3	4	5	Total#
lexical tones	1	85.6	7.1	1.5	5.3	0.5	2565
	2	6.4	84.8	1.4	6.7	0.7	2961
	3	3.7	10.3	79.3	6.5	0.2	2007
	4	3.4	4.8	1.9	89.3	0.7	4242
	5	1.8	1.5	0.4	1.4	95.0	1368

3) Relationship between Tone Value Classes and Tone Classes: By analyzing the co-occurrence between tone value classes and tone classes, the top-2 tone value classes for each of tones 1 to 5 are the same as summarized in Table III. The top-2 tone values match well with the common-agreed levels and contours of the five Mandarin tones. The T1 has the lowest entropy over the corresponding tone value labels, inferring that the speakers could pronounce T1 well. On the contrary, T3 has the highest entropy and both low-dipping tone value (21) and rising tone value (24), indicating that the speakers probably may not know how to process tone three Sandhi. T1 (with a standard tone value label of 55), the high-level tone, had an accuracy rate of 35.58%. Nevertheless, it was also produced as 44 with 33.75%. The rising tone T2 (35) had an accuracy rate of 19.76%, and was realized as 24 with 38.04%. This indicates that when L2 learners produce a rising tone, the ascent is not large enough, leading to a tonal deviation.

TABLE III: The top-2 tone value classes and entropies regarding 41 tone classes for each of lexical and perceived tones.

syllable tone	top-2 tone values	entropy
lexical T1	44 (34%), 55 (36%), others (30%)	1.874
lexical T2	24 (38%), 35 (20%), others (42%)	2.196
lexical T3	21 (38%), 24 (11%), others (51%)	2.374
lexical T4	53 (32%), 42 (28%), others (40%)	2.136
lexical T5	3 (26%), 2 (16%), others (58%)	2.339
perceived T1	44 (37%), 55 (39%), others (24%)	1.654
perceived T2	24 (40%), 35 (21%), others (39%)	2.084
perceived T3	21 (44%), 24 (10%), others (46%)	2.193
perceived T4	53 (37%), 42 (29%), others (34%)	1.937
perceived T5	3 (27%), 2 (17%), others (56%)	2.291

B. Speech Corpora For Pre-trained Model

Since the size of the target corpus (the Romance L2 Mandarin speech corpus) is small, we use the following three speech corpora to obtain a pre-trained tone recognition model for the following fine-tuning of the target tasks, i.e., tone class and tone value class recognitions for the target corpus:

1) AISHELL-1: AISHELL-1 [17] is an open-source Mandarin speech corpus for constructing Mandarin automatic speech recognition (ASR) systems. The corpus features recordings from 400 diverse Chinese speakers, high-quality audio at 16kHz, and over 95% transcription accuracy, ensured by expert annotation and thorough quality checks.

2) AISHELL-3: The AISHELL-3 [18] corpus is a collection of speech data with a total length of 85 hours with 997,998 syllables. It is mainly intended for constructing a text-to-speech system for multiple speakers. The database is labeled with high-accuracy phonetic and tone transcription, which makes it suitable for training and testing tone recognition models. In this study, we use the AISHELL-3 corpus to evaluate the performances of the pre-trained tone recognizers.

3) NER: The speech data of the NER corpus [19] originated from the National Education Radio Station in Taiwan. In total, the corpus contains about 5,129 hours and 6,658 files, of which 1,218 hours of data was transcribed by professional dictators, and the rest of the data was transcribed by ASR systems. We remove the utterances that can not be force-aligned successfully by the force-aligner used in the SLMTK [16] because we found that most utterances without successful speech alignments usually contain erroneous transcriptions. As a result, we used a total of 2,318 hours of the NER corpus to construct a pre-trained tone recognition model.

III. Pitch Analyses for Tone Values

A. Parameterized Raw Pitch Contour (RPC) Representation

This study employs the discrete orthogonal polynomials [20] widely used in modeling syllabic pitch contours of Mandarin [21]–[24] and Chinese dialects [25] to parameterize variable-length frame-based logF0 values (extracted by the RAPT [26]) with fixed-dimensional representation and small fitting errors. We can easily model the logF0 contours for some specific tone values and visualize them for illustration with this representation by

$$a_n(j) = \frac{1}{M_n+1}\sum_{m=0}^{M_n} F(n,m)\phi_j(m/M_n) \text{for } j = 0 \sim 3 \quad (1)$$

where $F(n,m)$ is the observed logF0 value at m-th voiced frame of the n-th syllable; $\phi_j(m/M_n)$ is the j-th orthogonal polynomial basis (j-th order polynomial) with a length of $M_n + 1$ voiced frames; $F(n,m)$ for $m = 0, 1, ...M_n$ is the so-called Raw Pitch Contour (RPC); $a_n(j)$ is called the orthogonal expansion coefficient (OEC) for j-th orthogonal polynomial basis, $\phi_j(m/M_n)$. The frame logF0 values can be reconstructed with a negligible fitting error by the four OECs, i.e., $a_n(j)$ for $j = 0, 1, 2, 3$ with:

$$\hat{F}(n,m) = \Sigma_{j=0}^{3} a_n(j)\phi_j(m/M_n) \text{ for } m = 0 \sim M_n \quad (2)$$

The four elements, i.e., $a_n(j)$ for $j = 0, 1, 2, 3$ in order represent respectively the mean, slope, acceleration and curvature of the logF0 contour.

B. Typical logF0 Patterns for Tones and Tone Values

Fig. 1a and Fig. 1b show average logF0 contours $\Sigma_{j=0}^{3}\beta_t(j)\phi_j(m/M)$, i.e., average RPCs, for the five lexical tones and perceived tones, respectively; where $\beta_t(j)$ is the average OEC for tone t calculated from all the syllable samples of the Romance corpus:

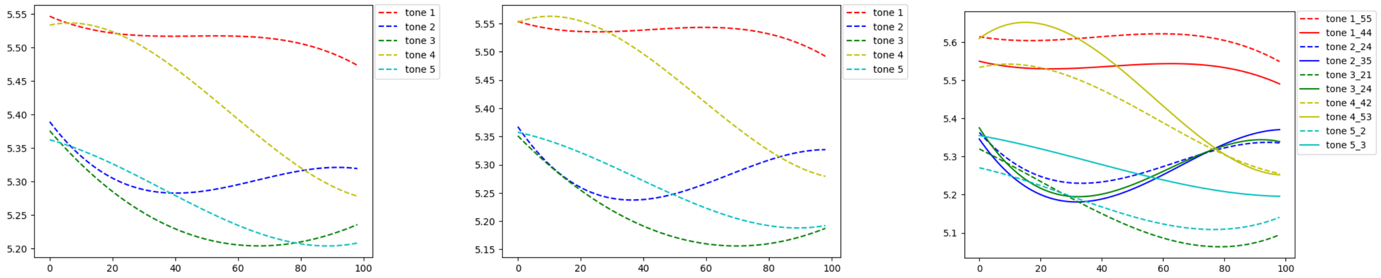
$$\beta_t(j) = (\Sigma_{n=1}^{N} a_n(j)\delta(t_n = t))/(\Sigma_{n=1}^{N}\delta(t_n = t)) \quad (3)$$

where t_n is the n-th syllable's lexical or perceived tone class.

It was found that the average logF0 for perceived T1 is more flat than the one for lexical T1. In addition, perceived T2 and T4 have more significant curvatures than lexical T2 and T4. We then calculate the mean square errors (MSEs) of the fitting for logF0 contour represented in OEC, a_n by the typical tone patterns, β_{t_n} , i.e., $\mathbb{E}[||a_n - \beta_{t_n}||_2^2]$. It is found that the MSE for the lexical tones (0.0761 ($\log \text{Hz}^2$)) is larger than the one for perceived tones (0.0704 ($\log \text{Hz}^2$)). This infers that the speakers may utter different or erroneous tones from lexical tones. The perceived tones labeled by humans could better represent tone classes of syllables.

Fig. 1c shows average RPCs for the ten significant tone values as shown in Table III paired with five lexical tones. For T1, tone value 55 has a higher logF0 level than tone value 44, matching the prior knowledge about tone values' pitch level. For T2, tone value 24 is not shown paralleled with tone value 35, manifesting a different logF0 from our prior knowledge about tone values 24 and 35. The two pitch contours of T2 are not significantly different at the beginning and end and do not conform to the expert's markings. The apparent difference in slope between the two contours may indicate that the acceleration of the pitch contour affects the perception of pitch by humans. T3 has two major tone values of 21 for low dipping and 24 for tone three Sandhi. For T4, tone value 53 shows a higher pitch beginning than tone value 42. The neutral tone (T5) with tone value 3 is higher in pitch than tone value 2. The level-four pitch in set ”24” is obviously not at the same height as level-four in set ”44”, and the same situation also occurs in level-five in sets ”35” and ”55”, which is probably due to coarticulation. The above findings indicate that the average logF0 contours (RPCs) of tone values in continuous speech do not generally match well with conventional 5-level tone values suggested by linguistics. Further investigation is required for future studies.

We also calculate the MSE of the fitting for logF0 contour by the typical tone value patterns as shown in Fig. 1c. The MSE for the tone value classes is 0.0565 $\log \text{Hz}^2$, which is smaller than the MSEs for the lexical and perceived tones. The reduction of the MSE infers that tone value could give a



(a) lexical tones

(b) perceived tones

(c) tone value classes

Fig. 1: Typical logF0 contours for the lexical tones (a), perceived tones (b), and tone values (c) calculated from the Romance L2 Mandarin speech corpus. Note that the legends in (c), i.e., tone_x_yy denotes lexical tone x and tone value yy

more sophisticated description of logF0 contours for describing non-native L2 Mandarin prosody.

IV. Automatic Tone Value Recognition

The Romance L2 Mandarin speech corpus is used to conduct tone value recognition experiments with a transformer-based recognition model, which is a stack of three modified transformer encoder layers. The architecture of the recognition model is the same as the one used in the previous tone labeling [15] and tone representation experiment [24], achieving a tone recognition accuracy of 87.48%, comparable to the related studied reported [27]. The input feature vector of the recognizer for each n-th syllable consisted of 81 10ms-interval Mel-spectrogram around the target syllable and information about syllable boundary and frame index. Since the Romance L2 Mandarin corpus is too small to construct a robust recognition model, we adopted tone recognition models trained with AISHELL-1, AISHELL-3, and NER corpora as pre-trained models. We then fine-tuned these pre-trained models with the Romanace L2 Mandarin corpus to obtain more robust models than the ones without the pre-training.

A. *Pre-trained Model*

The pre-trained models are trained with the corpora considered in Section II, validated, and evaluated with the validation and test sets of AISHELL-3, which are the same sets used in the previous study [15]. As shown in Table IV, increasing training data indeed enhances the model's accuracy in tone recognition. Although Model 3's overall accuracy is higher than Model 2's, Model 2 has a higher accuracy for T2. This is mainly because the AISHELL-1 and NER datasets were not manually corrected; the tone labels were taken directly from the lexicon without considering tone Sandhi in consecutive T3 syllables. To address this problem, we re-labeled the training data for Model 3 by the re-labeling strategy proposed in [15]. The overall accuracy of the re-labeled Model 3 is 90.7%, slightly higher than the original Model 3. Additionally, the accuracy of each tone is higher than that of Model 2. This demonstrates that the tone re-labeling method can effectively address the tone Sandhi issue in tone labeling.

TABLE IV: The accuracies of tone recognition experiments on the test set of AISHELL-3 with various training datasets.

	training data	syllable#	overall accuracy (%) acc. of T1/T2/../T5
Model 1 (baseline)	AISHELL-3	774,714	87.4 90.3/87.2/80.0/91.3/75.2
Model 2	AISHELL-1&3	2,728,112	89.4 91.0/88.7/83.9/93.1/79.4
Model 3	AISHELL-1&3+ NER	37,399,546	90.4 92.9/87.9/85.5/94.2/83.4
relabeled Model 3	AISHELL-1&3+ NER	37,399,546	90.7 92.6/89.6/85.2/94.2/83.8

B. *Tone and Tone Value Recognitions for the L2 Dataset*

We construct recognition models for multi-speaker tone and tone value recognition tasks on the Romance L2 Mandarin corpus. In both tasks, we use the utterances corresponding to the first short text in the corpus as training data (2084 syllables) and the rest of the utterances as testing data (2179 syllables). Table V shows the tone recognition results for the Romance L2 Mandarin dataset. Note that the L2 dataset is labeled with perceived tones by the three native Mandarin-speaking annotators. The tone labels made by each of the annotators are regarded as one set, resulting in three training and test sets corresponding to the three annotators. We, therefore, construct the three tone recognition models with the three training and test sets to evaluate performance.

The strict-sense accuracies were calculated by averaging the accuracies of three sets labeled by the three annotators. On the other hand, the non-strict sense accuracies considered a predicted tone correct if it is one of the tones in the three sets labeled by the three annotators. We found that the larger the dataset used to train the pre-trained model, the better the performance after fine-tuning. Contrary to expectations, the relabeled Model 3 did not achieve better results. We found the accuracy drop from Model 3 to relabeled Model 3 is due to the degradation of tone recognition accuracies for T4 and T3 syllables.

For the tone value recognizer, we changed the output layer dimension from 5 (five tones) to 41 (tone value classes labeled by the annotators). As shown in Table VI, similar to the tone recognition task, the relabeled model achieved lower performance both before and after fine-tuning. Note that the

TABLE V: The strict and non-strict sense tone recognition accuracies with various model settings: a) W/O represents that the models are trained with the Romance L2 dataset only, and b) Model 1/2/3 and relabeled Model 3 are fine-tuned from the pre-trained models reported in Subsection IV.A. Note that the numbers left to and right to the arrows are accuracies made by the models before and after fine-tuning to the L2 dataset, respectively.

pretrained model	strict sense acc.(%)	non-strict sense acc.(%)
W/O	47.2	53.6
Model 1	58.8→65.7	66.5→73.2
Model 2	59.9→67.3	67.7→74.5
Model 3	71.2→75.0	79.6→83.1
relabeled Model 3	70.7→74.7	78.9→82.5

TABLE VI: The strict and non-strict sense tone value recognition accuracies with various model settings: a) W/O represents that the models are trained with the Romance L2 dataset only, b) Model 1/2/3 and relabeled Model 3 are fine-tuned from the pre-trained models reported in Subsection IV.A, and c) XLS-R is a cascade of the pre-trained SSL model, XSL-R [28], with a downstream trainable DNN.

pretrained model	strict sense acc. (%)	non-strict sense acc. (%)
W/O	36.0	47.2
Model 1	46.6	62.4
Model 2	47.5	65.0
Model 3	47.4	66.1
relabeled Model 3	46.0	64.8
XLS-R	42.8	56.0

TABLE VII: The stripped-down confusion matrix of the nine major tone value classes by the fine-tuned Model 3 in the strict sense.

tone value	top-3 recognized tone value classes and percentages (%)
55	55 (71.1%), 44 (13.1%), 53 (8.1%), others (7.8%)
44	44 (42.4%), 55 (30.6%), 42 (10.2%), others (16.8%)
24	24 (69.3%), 21 (11.1%), 35 (5.7%), others (13.8%)
35	35 (57.4%), 24 (22.9%), 21 (6.8%), others (12.9%)
21	21 (63.6%), 24 (11.3%), 32 (8.9%), others (16.2%)
42	42 (56.2%), 53 (27.2%), 44 (4.4%), others (12.2%)
53	53 (84.6%), 42 (10.7%), 55 (1.8%), others (2.9%)
2	21 (42.1%), 2 (14.5%), 32 (10.1%), others (33.3%)
3	3 (31.1%), 21 (13.7%), 32 (13.7%), others (41.5%)

calculation of the strict and non-strict sense accuracies follows the definition for the tone recognition results reported in Table V. In addition to using the proposed pre-trained models, we also experimented with the upstream self-supervised learning (SSL) model, XLS-R [28], for the task of tone value recognition. We changed the input parameters from Mel-spectrograms to representations of the output of the ninth layer of XLS-R (the best-performing layer tested) and implemented it using a simple DNN model with 3 layers (1024x512x256). Although the result obtained is better than that using no pre-trained model, it is worse than those using the proposed pre-trained model.

Table VII shows the confusion matrix for tone value recognition. We found that although the overall accuracy for tone value recognition was only 47.4%, the accuracy rates for 6 of the 9 major tone value classes exceeded this number. This result is highly consistent with the distribution of training samples. For example, the tone value '53', which has the largest number of samples, achieved an accuracy of 84.62%, while the tone value '2', which has the smallest number of samples, had the worst performance, with an accuracy of only 14.47%. The tone value classes regarding level pitch contours, i.e., 44 and 55, are easily mutually confused. The tone value classes regarding falling pitch contours, i.e., 42 and 53, are also easily mutually confused. The rising-pitch tone value classes, i.e., 24 and 35, would be jointly confused with the standard T3 lower dipping tone value, i.e., 21, maybe due to tone Sandhi. The tone values corresponding to neutral tones, i.e., 2 and 3, are easily confused jointly with the tone value 21. These confusions between the tone value classes seem related to the similarities between the typical logF0 patterns of tone value classes obtained in Fig. 1c, suggesting that the tone value recognition result is reasonable.

V. Conclusions and Future Works

This study explored the relationship between tone and tone values, visualized the typical pitch contours of tone values, and attempted to train tone and tone value recognition models for the L2 Mandarin speakers. For tone recognition, fine-tuning can indeed improve accuracy, but the accuracy for the L2 speakers is much lower than that for the L1 speakers (in the AISHELL-3 test set). For tone value recognition, although the model trained using our proposed pre-trained model performs better than the model trained using the SSL upstream representations (XLS-R), the accuracy for most of the tone values is still not high. Based on the experimental results, we propose the following future research directions:

1) Analyze the samples of tone recognition errors in L1 and L2 corpora to understand the characteristics of recognition error samples for each tone class and the specific differences between L1 and L2 speakers in the pronunciation of tones.

2) Use the proposed tone value recognition model to label the unlabeled L2 corpus, conduct linguists analyze whether this method can accelerate their manual tone value annotation works, and improve the quality of Mandarin language teaching.

3) Add corpora from various languages to train the pre-trained model to improve the model's ability to analyze pitch information from the spectrogram, thereby enhancing the performance of the tone and tone value recognition model.

References

[1] W. Li, N. F. Chen, S. M. Siniscalchi, and C.-H. Lee, "Improving Mispronunciation Detection of Mandarin Tones for Non-Native Learners With Soft-Target Tone Labels and BLSTM-Based Deep Tone Models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2012–2024, Dec. 2019, ISSN: 2329-9290, 2329-9304.

[2] Y.-S. Cheng, T.-H. Lo, and B. Chen, "Exploiting text prompts for the development of an end-to-end computer-assisted pronunciation training system," in *Proceedings of the 32nd Conference on Computational Linguistics and Speech Processing (ROCLING 2020)*, 2020, pp. 290–303.

[3] Z. Gan, T. Zhao, X. Yu, and H. Yang, "Deep feedforward sequential memory networks based mispronunciation detection for tibetan students' mandarin," in *2021 2nd International Conference on Information Science and Education (ICISE-IE)*, IEEE, 2021, pp. 509–512.

[4] Y. Chao, *Tone, Intonation, Singsong, Chanting, Recitative, Tonal Composition, and Atonal Composition in Chinese*. 1956.

[5] Y. Chao, *A Grammar of Spoken Chinese*. University of California Press, 1968, ISBN: 978-0-520-00219-7.

[6] C. Zhu, *Pronounce Chinese as Chinese*. Beijing: Yuwen Publishing, 1997.

[7] C. Zhu, *Pronounce Chinese as Chinese (new edition)*. Taiwan: New Xue Lin Publishing Co., Ltd., 2013.

[8] C.Shih, "The phonetics of the chinese tonal system," AT T Bell Laboratories, Tech. Rep., 1986.

[9] J. A. Edmondson, J.-L. Chan, G. B. Seibert, and E. D. Ross, "The effect of right-brain damage on acoustical measures of affective prosody in Taiwanese patients," *Journal of Phonetics*, vol. 15, no. 3, pp. 219–233, 1987, ISSN: 0095-4470.

[10] J. Fon and W.-Y. Chiang, "What does chao have to say about tones?-a case study of taiwan mandarin," *Journal of Chinese Linguistics*, vol. 27, no. 1, pp. 13–37, 1999.

[11] P. Rose, "Considerations in the normalisation of the fundamental frequency of linguistic tone," *Speech communication*, vol. 6, no. 4, pp. 343–352, 1987.

[12] X. Zhu *et al.*, "Shanghai tonetics," 1995.

[13] J.-S. Jhang, *Let's study Thai language*, Xiu ding yi ban. Xinbei Shi: Uni-President Publishing House Co., Ltd., 2022, OCLC: 1373231644, ISBN: 978-986-99369-8-9.

[14] J.-M. Liang, *Learn basic Thai in 7 days(in chinese)*, first edition. Taipei: Wo-Shih publishing house, 2014, OCLC: 900451375, ISBN: 978-986-5785-42-0.

[15] W.-H. Li, C.-Y. Chiang, and T.-H. Liu, "Tone labeling by deep learning-based tone recognizer for mandarin speech," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2023, pp. 873–880.

[16] C.-Y. Chiang, W.-H. Li, Y.-T. Lin, *et al.*, "The speech labeling and modeling toolkit (slmtk) version 1.0," in *2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, 2022, pp. 1–5.

[17] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017, pp. 1–5.

[18] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A Multi-Speaker Mandarin TTS Corpus," in *Proc. Interspeech 2021*, 2021, pp. 2756–2760.

[19] Y.-F. Liao, Y.-H. S. Chang, Y.-C. Lin, W.-H. Hsu, M. Pleva, and J. Juhar, "Formosa Speech in the Wild Corpus for Improving Taiwanese Mandarin Speech-Enabled Human-Computer Interaction," en, *Journal of Signal Processing Systems*, vol. 92, no. 8, pp. 853–873, Aug. 2020, ISSN: 1939-8018, 1939-8115.

[20] S.-H. Chen and Y.-R. Wang, "Vector quantization of pitch information in Mandarin speech," en, *IEEE Transactions on Communications*, vol. 38, no. 9, pp. 1317–1320, Sep. 1990, 24, ISSN: 00906778.

[21] C.-Y. Chiang, S.-H. Chen, H.-M. Yu, and Y.-R. Wang, "Unsupervised joint prosody labeling and modeling for Mandarin speech," en, *The Journal of the Acoustical Society of America*, vol. 125, no. 2, pp. 1164–1183, Feb. 2009, ISSN: 0001-4966.

[22] S.-H. Chen, C.-H. Hsieh, C.-Y. Chiang, *et al.*, "Modeling of Speaking Rate Influences on Mandarin Speech Prosody and Its Application to Speaking Rate-controlled TTS," en, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 7, pp. 1158–1171, Jul. 2014, 09, ISSN: 2329-9290, 2329-9304.

[23] C.-H. Lin, C.-L. You, C.-Y. Chiang, Y.-R. Wang, and S.-H. Chen, "Hierarchical prosody modeling for Mandarin spontaneous speech," en, *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. 2576–2596, Apr. 2019, 41, ISSN: 0001-4966.

[24] W.-H. Li, T.-H. Liu, and C.-Y. Chiang, "Tone Value Representation for Computer-Assisted Pronunciation Training," in *Proc. Speech Prosody 2024*, 2024, pp. 712–716.

[25] C. Y. Chiang, "Cross-Dialect Adaptation Framework for Constructing Prosodic Models for Chinese Dialect Text-to-Speech Systems," en, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 108–121, Jan. 2018, 16, ISSN: 2329-9290, 2329-9304.

[26] D. Talkin and W. B. Kleijn, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995, 22.

[27] J. Tang and M. Li, "End-to-end mandarin tone classification with short term context information," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 878–883.

[28] A. Babu, C. Wang, A. Tjandra, *et al.*, *Xls-r: Self-supervised cross-lingual speech representation learning at scale*, 2021. arXiv: 2111.09296 [cs.CL].