SVM-based evaluation of Thai tone imitations by Thai-naïve Mandarin and Vietnamese speakers

Juqiang Chen*, Tianyi Ni†, Benjawan Kasisopa*, Mark Antoniou* and Catherine Best*
* Western Sydney University, The MARCS Institute for Brain Behaviour and Development, Sydney, Australia
E-mail: j.chen2@westernsydney.edu.au
† Department of Linguistics, The Ohio State University, Columbus OH, USA
E-mail: ni.386@osu.edu

***Abstract*—Native listener judgements and acoustic comparisons are sensitive to deviations between non-native speech and native productions, but both have drawbacks and are inefficient for evaluating large databases. To probe whether Support Vector Machines (SVM) might offer an efficient alternative, we used three SVM models trained with native Thai lexical tones to evaluate new native stimuli and non-native imitations by Mandarin and Vietnamese speakers. The optimal SVM model categorized native tones accurately but showed lower accuracy with non-native imitations, like native judges do, thus confirming its sensitivity to deviations from native productions. Thai falling tone imitations yielded the lowest classification accuracy, indicating that both groups' imitations were constrained by their native falling tones. Thai rising tones were better recognized for Vietnamese than Mandarin imitators, reflecting differences between their native rising tones. Thus, SVM modeling may provide an effective alternative to traditional perceptual- or acoustic-based evaluations of non-native speech.**

I. Introduction

Machine learning classification algorithms allow multiple acoustic correlates of speech to be modelled across languages. They have been used to investigate the contribution of acoustic features to classification of phonetic categories, using native speaker data to both train and test the models [1], [2]. Machine learning has also been used to characterize acoustic-phonetic similarities between languages. A model is trained to classify native phonetic categories of one language, then used to classify speech from another language in order to predict how its listeners will categorize the former language's phones into their native systems [3]–[6]. Here we offer a third use of machine learning in speech research: using models trained on native speech to evaluate non-native imitations.

Traditionally, human judgements or acoustic comparisons between native and non-native productions are used for such evaluations. For example, Wayland in [7] compared Thai lexical tones in target words produced by native speakers versus English learners of Thai, using both native listener ratings and acoustic analysis. Native listener ratings revealed which tone types were perceived to be strongly non-native accented, whereas acoustic analysis identified the acoustic properties that distinguish non-native productions from native productions. Native perceptual judgements are more holistic and easier to interpret than acoustic analyses, i.e., higher accuracy indicates more target-appropriate productions. However, a serious pragmatic constraint is that the large number of tokens generated by an L2 production/imitation study makes it quite challenging to get multiple perceptual evaluations of all tokens by each native judge [8]. Thus, human perceptual judgements are labor-intensive and raise issues of consistency within and across listeners, speakers and tokens.

In contrast to native perceptual judgements, analysis of acoustic properties is sufficient for phonetic categories that can be captured adequately with only one acoustic correlate, such as voice onset time for consonant voicing contrasts. Lexical tones, on the other hand, require multiple measures such as F0 mean, F0 onset and F0 offset values which must be evaluated via multiple formal models. This makes it difficult to evaluate overall good or poor performance compared with native productions, as native judges can do.

Given the complementary limitations of native perceptual judgments and previous comparisons of cross-language acoustic differences, it is desirable to develop more effective ways to simultaneously consider multiple acoustic dimensions and produce classification scores that are more analogous to native listener classifications. Machine learning models trained with native productions can learn the multiple acoustic characteristics of those native categories. If non-native productions closely resemble native productions, native-trained machine learning models should classify them with high accuracy. However, the more the productions deviate from the target items, the lower the classification accuracy should be. In addition, machine learning models can consistently identify a large amount of data more efficiently than, and without the attentional limitations of, human judges.

The present study employed machine learning algorithms to evaluate the non-native lexical tone imitation data obtained in [9], [10] and explore the effect of native language influence, memory load, stimulus talker variability and vowel variability on imitation. Native language influence was estimated from the same Mandarin and Vietnamese imitators' actual perceptual assimilations of Thai tones to their native tones in previous perceptual studies [11], [12]. We expected the machine learning model to detect variations in Thai tone imitation accuracy by the Mandarin and Vietnamese participants that were in line with the similarity/dissimilarity they had perceived between the Thai tone and their native tone categories.

The other factors in our imitation study, memory load and stimulus variability, are known to affect non-native tone per-

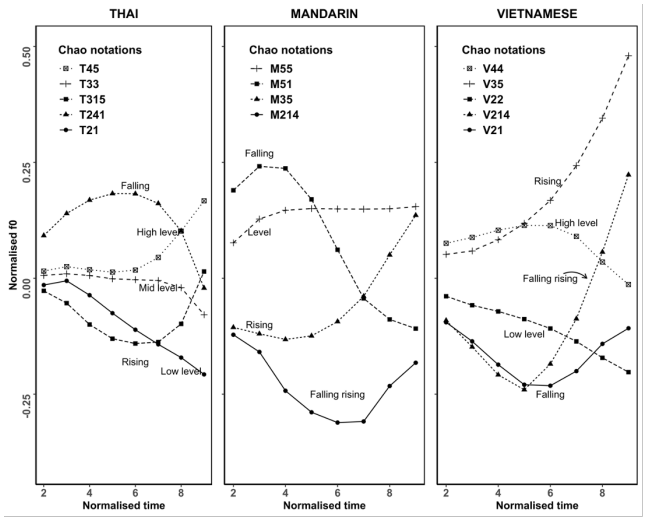

Fig. 1. Time- and Lobanov-normalised [21] F0 contours of Thai, Mandarin and Southern Vietnamese tones

ception [11], [12] and imitation [9]. The Automatic Selective Perception model [13] proposes that listeners attend less to phonetic details in a phonological than phonetic mode of perception. We reasoned that high memory load should bias listeners against a phonetic mode of perception because the full array of phonetic details decay in short-term memory [14], [15]. High memory load should instead bias them toward a phonological mode of perception in which they rely more on abstract phonological patterns and less on specific phonetic details [16], [17]. Because accurate imitation depends on retention of the phonetic details of the target, we expected non-native imitation to be less accurate and more constrained by native phonology under high than low memory load, and by high more than by low talker/context variability.

To provide a priori phonetic characterizations of the Thai, Mandarin and Vietnamese tones, we used Chao values [18], where F0 height at tone onset and offset, and sometimes an intervening value for a tone mid-point, are each referenced by the numbers 1-5 (low to high F0, respectively). Thai (T) has three level tones: high-level T45, mid-level T33, low-level T21, and two contour tones: rising T315 and falling T241 [19]. Mandarin (M) has four tones: level M55, rising M35, falling-rising M214 and falling M51 [18]. The dialect of our Southern Vietnamese participants has five tones: high-level V44, low-level V22, rising V35, falling V21, and falling-rising V214, [20] see Fig. 1.

II. Feature Selection for Tone Recognition

Lexical tones are acoustically represented by their F0 contours [22]. With the presence of a large number of features, a learning model tend to overfit, resulting in their performance degeneration when new test data is inserted [23]. Thus, based on previous phonetic research [24] that best represent the acoustic lexical tone data, for a given tone, an F0-related feature vector was used in our tone recognition scheme, which comprised the following 5 acoustic features:

- $F0_{onset}$: onset value of F0 contour
- $F0_{offset}$: offset value of F0 contour, which together with $F0_{onset}$ characterizes level, rising, or falling tones
- $F0_{mean}$: mean of F0 contour, which indicates tone height
- $F0_{excursion}$: the range of nominalized F0 contour, which distinguishes level tones from contour tones
- F0 maximum location or $F0_{max\_loc}$: the peak position relative to the tone's duration is measured, which allows us to distinguish differently timed peaks in convex and concave.

Other features, such as syllable duration or tone contours of the neighboring syllables, have been used in some tone recognition studies [25], [26]. However, we did not select F0 features of neighboring tones because our goal was to build models that classify monosyllabic Thai tones in isolation and evaluate non-native imitation of these tones in the same context. We did not aim to build an automatic speech recognition model to identify Thai running speech, which requires a larger quantity of training data from many phonetic environments. In addition, syllable duration was not selected because our focus was on tone imitation and not on syllable or vowel imitation. Given that vowel duration is phonologically contrastive in Thai but neither in Mandarin nor Vietnamese, duration measures could confound our results in evaluating F0 accuracy in tone imitations.

III. Tone Recognition Based on SVM

A. Multi-class support vector machine classifier

Thai tone recognition is essentially a multi-class classification problem, in which a tone category is assigned to a new data set with specific features. Given a set of $\mathcal{L}$ labeled examples, each data point has two parts: the d-dimensional vector of the acoustic features, with $d = 5$, and the corresponding labels of class:

$$S = \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathbb{R}^d, y_i \in 1, 2, 3, 4, 5, i = 1, ..., \mathcal{L}\} \quad (1)$$

where $\mathbf{x}_i$ is the vector of data features and, y_i is one of the five tone labels. The SVM maps the d-dimensional input vector $\mathbf{x}$ from the input space to the d_h-dimensional feature space with a non-linear function $\Phi(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^{d_h}$. The separating hyperplane in the feature space is defined by

$$\omega \cdot \Phi(\mathbf{x}) + b = 0, \ \omega \in \mathbb{R}^{d_h}, b \in \mathbb{R} \quad (2)$$

The classifier should satisfy the condition of existence of both ω and b such that

$$y_i(\omega^T \cdot \Phi(\mathbf{x}_i) + b) \geq 1 \quad (3)$$

Practically, the data from the five classes are not perfectly sparse, meaning that the data from neighboring classes in the hyperspace might overlap each other, which makes a perfect linear separation impossible. Hence, a restricted number of misclassifications are tolerated around the margins. The resulting optimization problem for SVM, in which the violation of the constraints is penalized, is given by:

$$\min_{\omega^m, b^m, \xi^m} \frac{1}{2}||\omega^m||^2 + C\sum_{i=1}^{\mathcal{L}} \xi_i^m \tag{4}$$

subject to

$$\begin{cases} y_i[(\omega^m)^T \cdot \Phi(\mathbf{x}_i) + b^m] \geq 1 - \xi_i^m & \text{if } y_i = m \\ y_i[(\omega^m)^T \cdot \Phi(\mathbf{x}_i) + b^m] \leq -1 + \xi_i^m & \text{if } y_i \neq m \end{cases} \tag{5}$$
$$\xi_i^m \geq 0, i = 1, ..., \mathcal{L}$$

where ξ_i is the relaxation factor tolerating misclassification, C is the penalty parameter controlling the tradeoff between allowing training errors and forcing strict margins, and m^{th} SVM is trained in the form of a one-against-the-rest approach (discussed in the next section). Generally, the constrained optimization problem is referred as the primal optimization problem, which can be written in the dual space by Lagrange multipliers $\alpha_i \geq 0$. The solution should maximize the following expression

$$\begin{aligned} L(\omega^m, b^m, \psi_i^m, \alpha_i^m) = \frac{1}{2}||\omega^m||^2 + C\sum_{i=1}^{\mathcal{L}} \xi_i^m \\ - \sum_{i=0}^{\mathcal{L}} \alpha_i^m [y_i((\omega^m)^T \cdot \Phi(\mathbf{x}_i) + b^m) - 1] \end{aligned} \tag{6}$$

The dual problem is given as

$$\max_{\alpha^m} L(\alpha^m) = \sum_{i=1}^{\mathcal{L}} \alpha_i^m - \frac{1}{2}\sum_{i,j=1}^{\mathcal{L}} \alpha_i^m \alpha_j^m y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{7}$$

subject to

$$0 \leq \alpha_i^m \leq C, \ \sum_{i=1}^{\mathcal{L}} \alpha_i^m y_i = 0 \tag{8}$$

The kernel function $K(x_i, x_j)$ corresponds to the inner product belonging to the transformation space:

$$K(x_i, x_j) = \Phi(x_i)\Phi(x_j) \tag{9}$$

Typically, kernel functions can be a radial basis function (RBF), a linear function or a polynomial function. After solving all the equations above, we arrive at k decision functions:

$$\begin{aligned} (\omega^1)^T \cdot \Phi(\mathbf{x}_i) + b^1 \\ \vdots \\ (\omega^k)^T \cdot \Phi(\mathbf{x}_i) + b^k \end{aligned} \tag{10}$$

The final classification function can be written as

$$\text{class}(\mathbf{x}_i) = \text{argmax}_{m=1,...,k}[(\omega^m)^T \cdot \Phi(\mathbf{x}_i) + b^m] \tag{11}$$

B. Multi-class classification approach

For the multi-class classification task, we consider it to be a set of binary classification problems. In this modeling study, we have followed the one-against-the-rest approach [27]. This method works by constructing k binary classifiers. The i^{th} classifier is trained using all positive-labeled examples. All other examples, regardless of their original valence, are then negative-labeled. The final output is the class that matches the classifier with the highest output value.

IV. Experiment Design

A. Data collection

Native Thai productions from a separate study [28] were used here with the authors' permission to build our machine learning models. 21 Native Thai speakers (13 female and 8 male speakers) were recorded in a sound-treated booth at Western Sydney University, using a Lavalier AKG C417 PP microphone at the sampling rate of 48 kHz and 16-bit resolution, as they produced multiple repetitions of each of the five Thai tones in citation form in five consonant-vowel syllables (/ma/, /mi/, /mu/, /na/, /ni/, /nu/ × 5 tones × 2-6 repetitions, 810 tokens in total).

The native Mandarin ($n = 32$) and the Vietnamese ($n = 32$) speakers from [12] also participated in the imitation experiment. Each language group was divided evenly into the low and high memory load conditions (Mandarin: low, M_{age} = 26.6 yrs, SD = 7 yrs, 10 females; high, $M_{age} = 26.0$ yrs, SD = 6.9 yrs, 10 females; Vietnamese: low, $M_{age} = 24.4$ yrs, SD = 7.7 yrs, 13 females; high, $M_{age} = 27.2$ yrs, SD = 12.8, 12 females), which differed in the interval between the offset of the stimulus and the signal to produce the imitation (low: 500 ms vs. high: 2000 ms). The Thai stimuli were presented in variable blocks with two talkers and/or two vowels and in constant blocks with just a single talker and vowel. The blocks were presented in random order.

Mandarin and Vietnamese participants were recorded individually in testing booths at Western Sydney University, University of New South Wales and Macquarie University. The target stimuli (five Thai tones each for the syllables /ma:/ and /mi:/) were presented from a Dell Latitude 7280 laptop running E-Prime via Sennheiser HD 280 pro headphones at 72 dB SPL. The imitations were recorded with a portable digital speech recorder (ZOOM H4n) with 41 kHz sampling rate and 16-bit stereo format. Participants were instructed to imitate the tones as faithfully as possible after they received the signal to respond. Each participant completed 160 imitation trials (5 tones × 2 syllables × 2 tokens × 2 talker variability × 2 vowel variability × 2 repeats) in total. Before the test session, they received 10 practice trials, with stimuli not used in the test.

All syllables were annotated and analyzed using the *Praat* script *ProsodyPro* [29], which was used to measure syllable duration and 10-equidistant points of F0 values (in Hz). The most stable part of the normalized tone (points 2 to 9) was used to calculate all F0-related measures. Those eight raw F0 values per imitation token were normalized by speaker

using the Lobanov method [21], which reflects how much an F0 value for a tone varies from the mean F0 of the speaker. This normalization process renders raw F0 values comparable among different speakers. As noted earlier, the acoustic measures we calculated for use in model training and evaluation of imitations were $F0_{onset}$, $F0_{offset}$, $F0_{mean}$, $F0_{excursion}$ and $F0_{max\_loc}$.

B. Training set sampling method

To train the SVM models with Thai native tones, native Thai tone dataset was shuffled randomly and partitioned into three disjoint sets of equal size, 2/3 set as training dataset and 1/3 set as testing dataset.

Given the small size of dataset and to give an unbiased estimate of the performance, the k-fold cross-validation method was also employed, in which every sample can be allocated in the testing set once and the variance of results is reduced as k increases [30]. The training algorithm must be rerun from scratch k times, but since we were applying SVM to a small dataset, the computation involved was modest. In our experiment, we performed a ten-fold cross-validation (CV).

C. Tone recognition and evaluation

The kernel functions RBF, linear and polynomial were each applied to work along with the SVM. The penalty parameter was set $C = 1.5$. The experimental results are shown in Table I. The 10-folded cross validation shows that the SVM model with the RBF kernel function showed the best performance in native Thai tone recognition, and the low variance tells us that the current experiment setting is unlikely to result in overfitting. Therefore, we chose that model to evaluate non-native imitation of tones by Mandarin and Vietnamese imitators. The overall average accuracies for the non-native imitation data are 66.5%, 70.5% and 69.6% for the SVM models with the polynomial, RBF and linear kernels respectively. Fig. 2 indicates variations in the performance of the RBF kernel SVM model over the five tone types for native Thai tone recognition. T315 was recognized with the lowest accuracy relative to other Thai tones. Fig. 3 indicates variations of the same model's performance over the five tone types for non-native imitation by Mandarin and Vietnamese imitators.

Assuming that this SVM model applied the same objective standard in identifying non-native tone categories as in identifying native tones, lower accuracies in recognizing non-native imitations reflect deviations of non-native imitations

TABLE I
ACCURACIES OF MODELS AND DATA

	POLYNOMIAL	RBF	LINEAR
Thai	88.8	94.7	87.2
10-folded CV	87.1±1.0	93.2±1.2	85.1±1.8
Mandarin	66.2	68.9	68.3
Vietnamese	66.7	72.1	70.9
Non-native overall	66.5	70.5	69.6

[a] Mean accuracies (%) of native-Thai-data-trained models and their accuracies in classifying Mandarin and Vietnamese data. 10-folded CV with means and standard deviations.

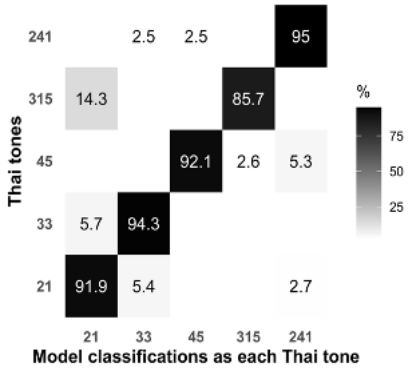

Fig. 2. Confusion matrices for native Thai tone productions as classified by RBF-based SVM model.

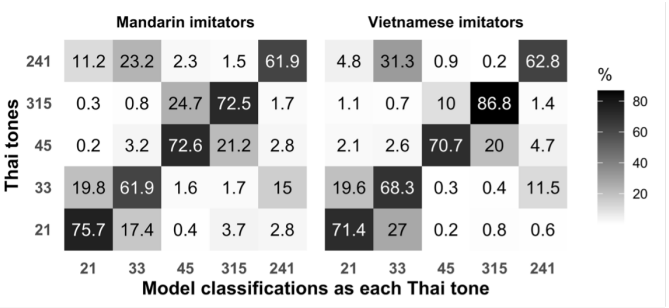

Fig. 3. Confusion matrices for Thai tone imitations by Mandarin (left) and Vietnamese imitators (right) as classified by the RBF-based SVM model.

from native productions. Thus, we calculated accuracy scores for each participant on each tone in each imitation condition as a measure of imitation performance. Then we carried out statistical analysis typical in psycholinguistic research to reveal the effects of native language, memory load and stimulus variability on non-native imitation. SVM identification accuracies were modeled as dependent variables for a linear mixed-effects model with participants as a random effect. Native language (Mandarin vs. Vietnamese) and memory load (low vs. high) were the between-subject fixed factors, whereas tone type (five Thai targets), target stimulus talker variability (constant vs. variable) and vowel variability (constant vs. variable) were the within-subject fixed factors. To calculate the p-values for the fixed effects, we used the Kenward-Roger approximation to degrees of freedom [31], and the Anova function from the car package in R, with test specified as "F". The significance level was set at .05.

There was only one main effect, i.e., tone type ($F_{(4,1112)} = 22.84$, $p < .001$) and three two-way interactions all involving tone type, i.e., language group $\times$ tone type ($F_{(4,1113)} = 6.15$, $p < .001$), memory load $\times$ tone type ($F_{(4,1113)} = 6.12$, $p < .001$), target stimulus talker variability $\times$ tone type ($F_{(4,1112)} = 20.68$, $p < .001$); and a three-way interaction language group $\times$ memory load $\times$ tone type ($F_{(4,1113)} = 2.78$, $p = .02$). We conducted multiple comparisons with Tukey adjustments to break down the main effect of tone

type (see Table II). T315 imitations ($M_{T315} = .81$) were classified significantly better than those of other Thai tones ($M_{T33} = .67$, $M_{T45} = .73$, $M_{T241} = .64$, $M_{T21} = .74$). T241 imitations were classified significantly worse than those of other Thai tones. T33 imitations were classified significantly *worse* than those of T21 and T45.

For all two-way and three-way interactions that involve tone types, only significant differences between other factors/levels for the same tone type are reported here. For the language group $\times$ tone type interaction, we found that Vietnamese imitations of T315 ($M = .87$) were better classified than Mandarin imitations ($M = .75$), $\beta = -.12$, $SE = .03$, $t(243) = -3.56$, $p = .01$.

For the memory load $\times$ tone type interaction, there were no differences between the two memory loads for the same tone type. For the target stimulus talker variability $\times$ tone type interaction, the classification accuracy for T241 imitations was lower in constant ($M = .52$) than variable ($M = .76$) talker blocks, $\beta = .24$, $SE = .03$, $t(1112) = 8.72$, $p < .001$. For the language group $\times$ memory load $\times$ tone type interaction, there were no differences between other factors/levels for the same tone type.

V. Discussion

First, the Thai-trained SVM model produced lower classification accuracy on non-native tone imitations than on native productions, indicating that it was sensitive to the acoustic deviations of the non-native imitations from native productions. T241 imitations yielded the lowest classification accuracy overall, across language groups. These same Mandarin and Vietnamese participants had shown a Categorized assimilation of T241 to their native falling tone M51 and level tone V44, respectively, which both differ in contour from T241 [11], [12]. Thus, the poor T241 imitations by the two groups can be attributed to the influence of their native tones.

Second, classification accuracies did not differ overall between the two language groups, suggesting that global differences between their native lexical tone systems in number and types of tones did not affect general ability to imitate non-native Thai tones. However, language backgrounds did interact with specific tone types, with multiple comparisons showing language group differences for imitations of the same Thai tone, T315. T315 was better recognized when imitated by the Vietnamese, who had perceptually assimilated it as clearly Categorized to V214, than by the Mandarin participants, who had instead perceptually assimilated it as Uncategorized and split between M35 and M214 [11], [12]. Strong native language influences in perceptual assimilation of particular non-native Thai tones thus appear to affect imitation of those tones, as we predicted.

TABLE II
Accuracies of models and data

Tones	β	SE	df	t	p
$21 - 33$	0.07	0.02	1113	3.56	0.004*
$21 - 45$	0.01	0.02	1112	0.57	0.980
$21 - 241$	0.10	0.02	1113	5.32	<.001*
$21 - 315$	-0.07	0.02	1111	-3.46	0.005*
$33 - 45$	-0.06	0.02	1112	-2.99	0.023*
$33 - 241$	0.03	0.02	1112	1.74	0.411
$33 - 315$	-0.13	0.02	1113	-7.00	<.001*
$45 - 241$	0.09	0.02	1114	4.74	<.001*
$45 - 315$	-0.08	0.02	1112	-4.02	0.001*
$241 - 315$	-0.17	0.02	1113	-8.76	<.001*

[a] Pairwise comparisons (with Tukey adjustments) among imitation tone types. Significant findings ($p < .05$) are marked with asterisk.

In addition, SVM model misidentification patterns for the imitations, as shown in Fig. 3, also indicate native language influences. For example, Mandarin imitations of T45 and T315 were most often misclassified by the Thai-trained models, which may be related to the Mandarin participants' perceptual assimilation of T45 as Categorized to M35, but also as assimilated to M214 with a lower percent choice [11]. On the other hand, their assimilation of T315 was Uncategorized, split more evenly between the very same two Mandarin tones, M35 and M214. This overlap in the Mandarin participants' assimilations of both T45 and T315 to the same two native tones appears to have affected their imitations of the two Thai tones. In addition, their previous perceptual assimilations of T21, T33 and T241 had also overlapped for two Mandarin response categories, M55 and M51, consistent with the SVM model's error response patterns to the imitations of these Thai tones, as shown in Fig. 3.

For Vietnamese participants, T45 and T315 had also overlapped in their perceptual assimilation to the native tone V214 [11], [12]. This may similarly explain why Vietnamese imitations of these two tones were confused by the SVM model. In addition, these participants had assimilated T241 and T33 with overlap in their native tones V44 and V22, and had assimilated T21 and T33 with overlap in V22. Thus, Vietnamese imitations of T241, T21 and T33 were treated by the machine learning model as more similar to each other than any of them were to T45 and T315.

Third, target stimulus talker variability, but not vowel variability, affected the SVM classification of the relevant imitations. However, imitation of T241 was acoustically more accurate in variable than constant talker blocks, the opposite pattern from our expectations. A possible psycholinguistic explanation for this discrepancy is that T241 is phonetically rising-falling and does not have phonetically similar counterparts in either Mandarin or Vietnamese. Thus, the phonetic details of T241 are both more complex and more unfamiliar as a phonetic contour to both non-native groups than other tones are. We speculate that in variable talker blocks, the varying phonetic details of these tokens forced the listeners to abstract the final falling contour from the T241 stimuli, which is more familiar as it is analogous to their native falling tones. On the other hand, in constant talker blocks, the imitators could focus on the specific phonetic details of each T241 stimulus token including the initial rise. Thus, they seem to have abstracted the general final falling contour of T241 in variable talker blocks, but to have gotten lost in the unfamiliar phonetic details of the T241 contour in constant talker blocks.

In conclusion, SVM models do appear to provide an efficient alternative approach to human perceptual judgements and acoustic comparisons for evaluation of non-native productions of lexical tones, and thus could provide feedback to learners. Classification results of these models are consistent with expected native language influences on non-native tone imitation, as interpreted in light of the perceptual assimilation patterns for those same tones by the same participants in our earlier studies. Additionally, talker variability can bias listeners to attend to phonological, i.e., more abstract, or phonetic, i.e., more detailed, levels of information, which in turn affects their non-native imitations. Machine learning models based on multiple dimensions of phonetically-relevant acoustic information could also be extended to evaluate non-native imitations of consonant and vowels in future research.

ACKNOWLEDGMENT

This research was supported by a China Scholarship Council and Western Sydney University Joint scholarship awarded to the first author, Juqiang Chen.

REFERENCES

[1] S. Li, W. Gu, L. Liu, and P. Tang, "The Role of Voice Quality in Mandarin Sarcastic Speech: An Acoustic and Electroglottographic Study," *J Speech Lang Hear Res*, vol. 63, no. 8, pp. 2578–2588, Aug. 2020, doi: 10.1044/2020_JSLHR-19-00166.

[2] W. Styler, "On the Acoustical and Perceptual Features of Vowel Nasality," PhD Thesis, University of Colorado at Boulder, 2015.

[3] J. Chen, C. T. Best, and M. Antoniou, "Predicting potential difficulties in second language lexical tone learning with Support Vector Machine models," presented at the ICWL 2020 and SETE 2020, China, 2020.

[4] P. Escudero and P. Vasiliev, "Cross-language acoustic similarity predicts perceptual assimilation of Canadian English and Canadian French vowels," *The Journal of the Acoustical Society of America*, vol. 130, no. 5, pp. EL277–EL283, Oct. 2011, doi: 10.1121/1.3632043.

[5] W. Strange, O.-S. Bohn, S. A. Trent, and K. Nishi, "Acoustic and perceptual similarity of North German and American English vowels," *The Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1791–1807, Mar. 2004, doi: 10.1121/1.1687832.

[6] W. Strange, O.-S. Bohn, K. Nishi, and S. A. Trent, "Contextual variation in the acoustic and perceptual similarity of North German and American English vowels," *The Journal of the Acoustical Society of America*, vol. 118, no. 3, pp. 1751–1762, Sep. 2005, doi: 10.1121/1.1992688.

[7] R. Wayland, "Non-native Production of Thai: Acoustic Measurements and Accentedness Ratings," *Appl Linguist*, vol. 18, no. 3, pp. 345–373, Sep. 1997, doi: 10.1093/applin/18.3.345.

[8] Y. Wang, A. Jongman, and J. A. Sereno, "Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training," *The Journal of the Acoustical Society of America*, vol. 113, no. 2, pp. 1033–1043, Jan. 2003, doi: 10.1121/1.1531176.

[9] J. Chen, C. Best, and M. Antoniou, "Cognitive Factors in Thai-Naïve Mandarin Speakers' Imitation of Thai Lexical Tones," in *Proc. Interspeech 2019*, 2019, pp. 2653–2657. doi: 10.21437/Interspeech.2019-1403.

[10] J. Chen, C. T. Best, and M. Antoniou, "Cognitive factors in Thai-naïve Mandarin and Vietnamese speakers' imitation of Thai lexical tones".

[11] J. Chen, C. Best, M. Antoniou, and B. Kasisopa, "Cognitive factors in perception of Thai tones by naïve Mandarin listeners.," in *Proceedings of the 19th ICPhS*, Canberra, Australia, 2019, pp. 1684–1688.

[12] J. Chen, C. Best, and M. Antoniou, "Phonological and phonetic contributions to perception of non-native lexical tones by tone language listeners: Effects of memory load and stimulus variability," In prep.

[13] W. Strange, "Automatic selective perception (ASP) of first and second language speech: A working model," *Journal of Phonetics*, vol. 39, no. 4, pp. 456–466, Oct. 2011, doi: 10.1016/j.wocn.2010.09.001.

[14] Y. Asano, "Discriminating Non-Native Segmental Length Contrasts Under Increased Task Demands," *Lang Speech*, vol. 61, no. 3, pp. 409–429, Oct. 2017, doi: 10.1177/0023830917731907.

[15] J. F. Werker and J. S. Logan, "Cross-Language Evidence for Three Factors in Speech Perception," *Perception Psychophysics*, vol. 37, no. 1, pp. 35–44, 1985, doi: 10.3758/BF03207136.

[16] J. S. Magnuson and H. C. Nusbaum, "Acoustic differences, listener expectations, and the perceptual accommodation of talker variability," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 33, no. 2, pp. 391–409, 2007, doi: 10.1037/0096-1523.33.2.391.

[17] J. A. Shaw and M. D. Tyler, "Effects of vowel coproduction on the timecourse of tone recognition," *The Journal of the Acoustical Society of America*, vol. 147, no. 4, pp. 2511–2524, Apr. 2020, doi: 10.1121/10.0001103.

[18] Chao. Y.R., "A system of tone-letters," *Le Maitre Phonetique*, vol. 45, pp. 24–27, 1930.

[19] A. Reid *et al*., "Perceptual assimilation of lexical tone: The roles of language experience and visual information," *Atten. Percept. Psychophys.*, vol. 77, no. 2, pp. 571–591, Feb. 2015, doi: 10.3758/s13414-014-0791-3.

[20] J. Chen, C. T. Best, and M. Antoniou, "Native phonological and phonetic influences in perceptual assimilation of monosyllabic Thai lexical tones by Mandarin and Vietnamese listeners," *Journal of Phonetics*, vol. 83, p. 101013, Nov. 2020, doi: 10.1016/j.wocn.2020.101013.

[21] B. M. Lobanov, "Classification of Russian Vowels Spoken by Different Speakers," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 606–608, Feb. 1971, doi: 10.1121/1.1912396.

[22] S. Wang, Z. Tang, Y. Zhao, and S. Ji, "Tone Recognition of Continuous Mandarin Speech Based on Binary-Class SVMs," in *First International Conference on Information Science and Engineering*, 2009, pp. 710–713. doi: 10.1109/ICISE.2009.1313.

[23] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review. Data classification: Algorithms and applications," in *Data Classification: Algorithms and Applications*, C. C. Aggarwal, Ed. 2015.

[24] J. Kuang, "The Tonal Space of Contrastive Five Level Tones," *PHO*, vol. 70, no. 1–2, pp. 1–23, 2013, doi: 10.1159/000353853.

[25] 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)*, 2012, pp. 4741–4744. doi: 10.1109/ICASSP.2012.6288978.

[26] H. Chao, C. Song, B.-Y. Lu, and Y.-L. Liu, "Feature Extraction based on DBN-SVM for Tone Recognition," *Journal of Information Processing Systems*, vol. 15, no. 1, pp. 91–99, Feb. 2019, doi: 10.3745/JIPS.04.0101.

[27] B. Aisen, "A Comparison of Multiclass SVM Methods," 2006. https://courses.media.mit.edu/2006fall/mas622j/Projects/aisen-project/index.html

[28] D. Burnham, T. Kuratate, C. McBride-Chang, and K. Mattock, "Making speech three-dimensional: Adding tone to consonant- and vowel-based speech perception and language acquisition research, quantification and theory." 2009. [Online]. Available: http://purl.org/au-research/grants/arc/DP0988201.

[29] Y. Xu, "ProsodyPro—A tool for large-scale systematic prosody analysis," 2013.

[30] J. Schneider and A. W. Moore, "A Locally Weighted Learning Tutorial using Vizier 1.0." 1997.

[31] U. Halekoh and S. Hojsgaard, "A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models–the R package pbkrtest," *Journal of Statistical Software*, vol. 59, no. 9, pp. 1–30, 2014.