Interspeech 2018
2-6 September 2018, Hyderabad

Analysis of L2 learners' progress of distinguishing Mandarin Tone 2 and Tone 3

Yue Sun[1], *Win Thuzar Kyaw*[1], *Jinsong Zhang*[2], *Yoshinori Sagisaka*[1]

[1]Graduate School of Fundamental Science and Engineering, Waseda University, Japan
[2]College of Information Science, Beijing Language and Culture University, China

yue.cherry.sun@gmail.com, winthuzarkyaw@akane.waseda.jp, jinsong.zhang@blcu.edu.cn, ysagisaka@gmail.com

Abstract

Many studies have shown the effectiveness of perceptual training to improve L2 learners' ability to distinguish Mandarin tones. In this paper, we quantified learners' perceptual characteristics on discriminating the most difficult tone pair, Tone 2 and 3 in Mandarin before and after training. L2 learners' categorical perception is measured by fitting a sigmoid curve to the identification responses with average F0 height as the acoustic dimension. The boundary location of the two tones in L2 learners' perception space is significantly improved to a higher F0 height after training. Regression analysis indicated that $\delta F0$ and δt of the initial fall of the concave F0 shape are the key acoustic features for native speakers in discrimination. L2 learners rely on not only the initial fall but also the $\delta F0$ of the final rise to discriminate the tones. A detailed analysis using cognitive measurements reports an increasing attention on the initial fall of the F0 contour for L2 learners after perceptual training. These results confirmed that directing the attention to key acoustic features is essential for L2 learners to improve their categorical perception of novel speech contrasts.

Index Terms: perceptual training, Mandarin tone, Japanese L2 learners, key acoustic feature

1. Introduction

It is well known that when learning Mandarin Chinese distinguishing the four lexical tones is a very difficult task, especially for learners who's native language is non tonal. In particular, the discrimination of the second and third tone (Tone 2 and 3) has been one of the toughest problems. According to acoustic analysis, this difficulty seems to result from the similar F0 contours of Tone 2 and 3 [1]. The citation form of Tone 3 in monosyllabic context is a falling-rising pitch. Although Tone 2 can be phonologically defined as a raising tone, there exists an obvious short F0 decrease following the onset. Thus, both Tone 2 and 3 are realized in a concave F0 shape [2][3].

There are major studies investigated distinctive acoustic features for native Mandarin speakers to discriminate Tone 2 and 3. Studies using multidimensional scaling models have demonstrated that F0 height and direction are the two fundamental dimensions to characterize the perceptual space of speakers of a tonal language [4]. In this two-dimensional perceptual space, Mandarin Tone 2 is positioned at a positive slope with middle-ranged height area while Mandarin Tone 3 is positioned at a middle-ranged slope with low height area [5][6]. Instead of using schematic features, studies focusing on explicit F0 shape have reported specific changes of the F0 shapes as key features in native speakers' discrimination. The turning point (the point at which the direction of the F0 contour changes from falling to rising) and the initial fall (the difference in F0 between onset and turning point) have been reported to be the main and the complementary cue [3] or vice versa [2]. The turning point is also demonstrated to be a complementary cue to pitch height [1]. Other than using F0 contour as distinctive features, it was found that amplitude contour is a perceptual cue for native speakers in discrimination when the F0 contour is absent [7], syllable duration is not a feature for distinguishing the two tones in perception [7], and, finally, creaky voice is a great feature for a tone being perceived as Tone 3.

Although these studies do not propose the same distinctive features to discriminate Tone 2 and 3 for native Mandarin speakers, it is agreed upon that the two tones are perceived categorically. On the contrary, L2 learners have difficulties to categorically discriminate the two tones [1]. Laboratory training is widely used to improve L2 learners' ability to distinguish the new phonetic contrasts. Their effectiveness to train L2 learners to perceive Mandarin tones has been shown. L2 learners gained a rapid increase of the identification accuracy with more robust categorical perception, as well as a great improvement of the generation ability [8][9]. Moreover, according to the results reported on individual variability in lexical tone learning for English-speaking learners of Mandarin, in the dimension of pitch height and pitch direction, perceptual training increased the learners' ability to identify tones by pitch direction [5].

In our previous perceptual training studies, Japanese L2 learners obtained a improvement of identification accuracy to distinguish Mandarin Tone 2 and 3 from 84% before training to 94% after training [10]. In the current study, the improvement of L2 learners' categorical perception after finishing the perceptual training is taken into consideration. Besides the increasing accuracy of identification, we investigate the goodness of the tonal categories in the L2 learners' perceptual space and their selective attention towards distinctive acoustic features to discriminate the two tones. The purpose of this analysis is to better understand the learning progress of L2 learners when categorical perception for novel speech contrasts is established.

In the following Sections, first, we explain the synthesized fundamental frequency continua and the two-step perceptual training method we used in the perceptual training experiment in Section 2. To confirm the improvement of L2 learners' perceptual ability, we measure the location and steepness of the participants' perceptual boundary in Section 3. In order to investigate L2 learners' selective attention of distinctive acoustic features to discriminate the tones, we use regression analysis to test the significance of F0 shape features for the participants identification responses, and observe the correlation with the cognition measurement and reaction time in Section 4. Finally, we conclude the analysis in Section 5.

10.21437/Interspeech.2018-1983

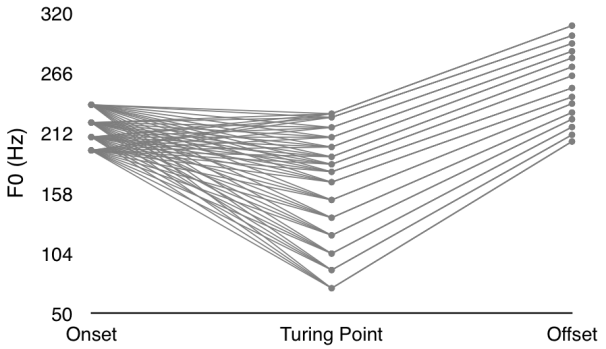

Figure 1: *Schematic diagram of the F0 contours of the synthesized Tone 2-3 continua.*

2. Data collection

2.1. Synthetic fundamental frequency continua

For training Japanese L2 learners to perceive the delicate changes of F0 shape between Tone 2 and 3, we manipulated the F0 height of the key points on the F0 contour, the onset, turning point and offset of the original speech sample to synthesize the Tone 2-3 continua. The segmental context was fixed to /da/ in the continua. The F0 range of the continua was defined by the speech range of the speaker who pronounced the original speech sample. By measuring the three key points' F0 height of 20 Tone 2 and 3 speech samples in total of the same speaker, the average F0 height plus one standard deviation of her T2s was set as the ceiling of the continua and the average F0 height minus one standard deviation of her T3s was set as the bottom of the continua. The F0 range was separated into 4 levels for the onset and 15 levels for the turning point and the offset. 15 stimuli for each continuum were created by manipulating the F0 height of the turning point and the offset at the same time but keeping the position of the turning point fix. Matching the 4 F0 levels of the onset with the 15 changes of the turning point and offset, we created 4 continua with the same F0 changes of the final rise but different F0 changes of the initial fall (Figure 1). In the end, we manipulated the syllable duration of the 4 continua from 320ms to 420ms, created 120 stimuli ($15 \times 4 \times 2$) in total.

2.2. Two-step perceptual training experiment

According to the hypothesis proposed by the popular training methods, we had designed a two-step perceptual training experiment using both synthesized and natural speech to train Japanese L2 learners to discriminate Mandarin Tone 2 and 3. The synthesized stimuli were used in the first training step in order to guide the learners attention towards the delicate changes of the F0 contour [11]. In the second training step, we employed natural speech samples of many different phonetic contexts by multiple speakers to get learners accustomed to different F0 shapes in rich speech contexts [12][13]. Participants of the perceptual training experiment were asked to spend 50 minutes per day on practicing with the speech samples. There were 2 days for the first training step and 4 days for the second training step.

A test session was conducted in order to collect the perceptual responses of the participants. It included the 120 synthetic stimuli and another 60 natural speech samples. The participants had to complete the test session three times during the training experiment, once before the first training session (pretest), once in between the first and the second training sessions (middle-test) and once after the second training session (post-test). In the test sections, the participants were required to identify whether they were listening to Tone 2 or 3 for each single stimulus they heard and indicate the answers by pressing the corresponding buttons on the keyboard (two-alternative-identification task). The speech samples were presented in random sequence by E-prime 2 [14]. The test started with 6 warming-up speech samples and there was a 3s interval in between the participants' response and presenting the next stimulus.

2.3. Participants

Participants of the perceptual training experiment were 13 Japanese native-speaking learners of Mandarin Chinese (age mean: 21, 5 females, 8 males). All of them had been studying Mandarin for less than 1 year in China at the moment of joining the experiment. Including the three tests and the two training sessions, each participant underwent a total of 5 experimental sessions. In order to compare the learners' perceptual patterns with Mandarin native speakers, 13 Mandarin native-speaking university students (age mean: 22, 7 females, 6 males) were recruited to complete the test section one time. None of the participants reported history of neurological or hearing deficits.

3. Categorical perception of Tone 2 and 3

In the present study, we analyzed the participants' response data to the synthetic continua.

3.1. Sigmoid curve fitting to the identification responses

A logistic regression model was fitted to the participants' proportion of Tone 2 responses of each continuum. The models included a bias coefficient (α), and an average F0-tuned coefficient (β_{aveF0}) given by the average F0 properties of the stimuli (x_{aveF0}).

$$ln\left(\frac{P_{(T2|x_{aveF0})}}{1-P_{(T2|x_{aveF0})}}\right) = \alpha + \beta_{aveF0} \times x_{aveF0} \quad (1)$$

Parameters of perceptual boundary were extracted from the fitted model to evaluate the participants' categorical perception [15][16]. (1) Boundary location of the two perceptual categories is the correspondence average F0 value when the predicted probability of Tone 2 is equal to 0.5. (2) Boundary steepness is the slope of the steepest tangent to the sigmoid curve, which is equal to one-quarter of the β_{aveF0}. The model fitting and statistic analysis were performed using the statistical programming language R [17].

3.2. Categorical perception of Tone 2 and 3

Figure 2 shows the distribution of the perceptual boundaries of the 8 synthetic continua for both natives and learners. The "pre", "mid", "post" and "native" hereafter refer to the pretest, middle-test and post-test for the learners and the test for the natives, respectively.

Comparing to the natives categorical perception, the boundary locations of the learners' perception are at a much lower F0 range regardless of the tests, which indicates a wider category range for Tone 2 but a narrower range for Tone 3. Natives' boundary steepness shows a much bigger value than the ones for the learners, which indicates a much steeper change from Tone

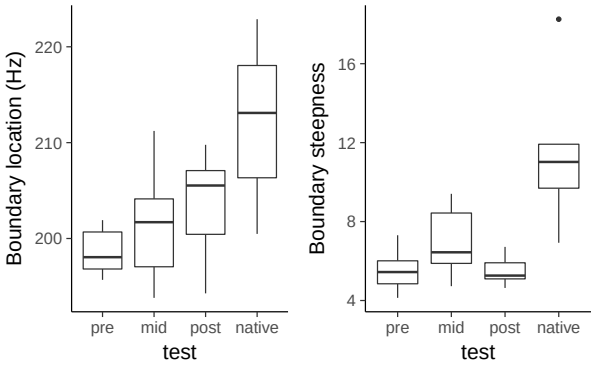

Figure 2: *Perceptual boundary between Tone 2 and 3 for the Japanese L2 learners of pretest (pre), middle-test (mid) and post-test (post) in the perceptual training experiment, and for the Mandarin-native (native) participants.*

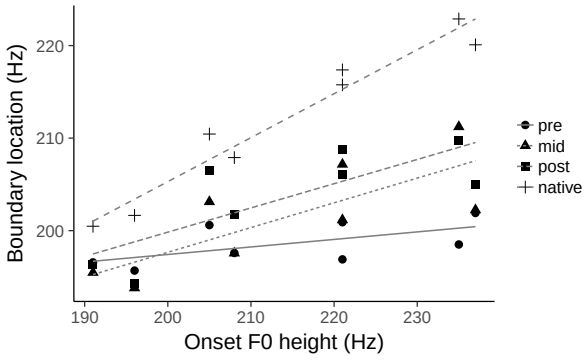

Figure 3: *Relationship between F0 height of the onset and the predicted boundary location of the synthetic continua for the native Mandarin speakers and the Japanese L2 learners.*

3 category to Tone 2 category, i.e. more robust tone categories for the natives. The perceptual training experiment induced a relatively higher F0 value for the perceptual boundary, indicating the extension of the T3 category and the shrinking of the T2 category. However, although the learners' boundary steepness became relatively steeper in the middle-test, it returned to where it was before training in the post-test.

A Two Sample t-test is used to compare the significance of differences between the natives and the learners as well as across the three tests for the learners. The differences are significant between natives and learners regardless of the tests, both for the boundary location and the steepness. The only significant difference for the learners across the three tests is the boundary location between the pretest and the post-test ($t(14) = 2.288, p = 0.038$).

3.3. Effects of F0 drop on boundary location

Figure 3 shows the relationship between the perceptual boundary location and the onset F0 height of the corresponding continuum for the natives and the learners. Result of the Pearson's correlation analysis shows that the boundary location shifts to higher F0 with the increase of the onset F0 height of the continuum for the natives ($r(6) = 0.97, p < 0.001$). For the learners, the relation is getting stronger in the middle-test ($r(6) = 0.78, p = 0.023$) and the post-test ($r(6) = 0.79, p = 0.020$) compared to the pretest ($r(6) = 0.60, p = 0.112$).

These results clearly show the importance of the F0 differences of the initial fall of the contour to the natives' boundary location in the perceptual space between the Tone 2 and 3 categories. The effect of the perceptual training method could also be seen through the correlation changes across the three tests for the learners. The training session with the synthetic stimuli successfully draws the learners attention to the F0 drop of the contour, then the training session with the natural speech samples pushes the learners' perceptual boundary to a higher F0 range.

Therefore, in order to better interpret the differences on categorizing the two tones between native Mandarin speakers and L2 learners as well as the learners perceptual progress during the perceptual training, we investigate the importance of F0 shape for the participants' categorical perception.

4. F0 shape as the key feature for discrimination

4.1. F0 drop as the key feature for native speakers

As the variation of F0 contour of Tone 2 and 3 is caused by the variation of F0 height and duration properties, we adopt 5 acoustic features to define the F0 shape. Taking the F0 height of the onset ($F0_{onset}$) as the reference F0 value, let $\delta F0_{drop}$ and $\delta F0_{rise}$ be the F0 differences of the initial fall and the final rise of the F0 contour, and δt_{drop} and δt_{rise} be the length of the initial fall and the final rise of the F0 contour.

The logistic regression model included a bias coefficient (α) and coefficients given by the 5 acoustic features were fitted to participants' proportion of Tone 2 responses of the 120 synthetic stimuli all in one time. The 5 acoustic features are examined by their significance on model fitting by a stepwise regression comparison task, in which the bidirectional elimination and Akaike information criterion (AIC) is used to compare the models. Coefficient values of the five predictors as well as the p-values are listed in Table 1.

The significant acoustic features of the natives' model are $\delta F0_{drop}$ and δt_{drop}. Neither $\delta F0_{rise}$ nor δt_{rise} has reached to a significant level. On the other hand, except $\delta F0_{drop}$ and δt_{drop}, $\delta F0_{rise}$ is significant in the learners' models for the three test. Furthermore, although $F0_{onset}$ is the reference value in the model fitting, it is not significant in the natives' model but very significant in the learners' models for the three tests. The δt_{rise} is not significant in neither the natives' model, nor the learners' models, which makes it a redundant feature for categorizing the two tones.

Since the models for the learners did not report change of the significance of any acoustic features, we could say that F0 differences of the final rise and the F0 height of the onset still play a very important role on learners' discrimination of Tone 2 and 3 even after finishing the perceptual training.

4.2. Better attention on the F0 drop for L2 learners after perceptual training

The reaction time for completing the identification task is used to analyze the participants' delicate perceptual changes during the perceptual training. Reaction time is a well used psychological parameter in multidimensional scaling analysis for measuring the perceptual distance between two auditory objects in discrimination tasks, such that the greater the reaction time the

Table 1: *Out put of the logistic regression with the five predictors for Mandarin natives and L2 learners across the three tests. P-value of each predictor is attached under the coefficient value in the brackets.*

	native	**pre**	**mid**	**post**
α	-34.66 ($p = 0.32$)	-201.8 ($p = 0.00$)	-158.3 ($p = 0.00$)	-144 ($p = 0.00$)
$\beta_{F0_{onset}}$	8.15 ($p = 0.17$)	36.68 ($p = 0.00$)	29.34 ($p = 0.00$)	26.5 ($p = 0.00$)
$\beta_{\delta F0_{drop}}$	17.69 ($p = 0.00$)	26.93 ($p = 0.00$)	24.26 ($p = 0.00$)	20.77 ($p = 0.00$)
$\beta_{\delta t_{drop}}$	-0.05 ($p = 0.01$)	0.01 ($p = 0.03$)	-0.02 ($p = 0.03$)	-0.01 ($p = 0.01$)
$\beta_{\delta F0_{rise}}$	-4.99 ($p = 0.51$)	27.3 ($p = 0.00$)	22.02 ($p = 0.00$)	19.83 ($p = 0.00$)
$\beta_{\delta t_{rise}}$	0.01 ($p = 0.16$)	-0.00 ($p = 0.97$)	0.00 ($p = 0.72$)	0.00 ($p = 0.51$)

closer in perceptual space are the two stimuli [18]. In our previous observation of two-alternative-choice identification tasks, participants' reaction time increased along with the proportion of their responses in one category decreased [19]. Thus, we speculate that the reaction time is highly related to the participants' proficiency of using key acoustic features to identify speech as a phonetic category.

To prove our hypothesis, we fit the logistic regression models again by using the $F0_{onset}$ and the key features of F0 drop that we suspect to be responsible for discriminating Tone 2 and 3, $\delta F0_{drop}$ and δt_{drop}, for the participants' responses, then compare the relationship between the reaction time and the predicted probability estimated by the models of each tone category. As expected, there is a strong negative correlation between the reaction time and the predicted probability of each tone category for the natives ($r_{T2}(54) = -0.81, p = 0.00$; $r_{T3}(62) = -0.71, p = 0.00$). It confirms the importance of F0 drop on natives' discrimination of Tone 2 and 3. That is, with the acoustic altering along F0 drop dimensions, the higher the probability a stimulus being perceived as a Tone 2 or 3, the shorter time it takes for the participant to respond. For the L2 learners, the correlation coefficients become more and more negative with the progress of the perceptual training experiment from the pretest ($r_{T2}(73) = -0.43, p = 0.00$; $r_{T3}(43) = -0.33, p = 0.03$) to the middle-test ($r_{T2}(69) = -0.62, p = 0.00$; $r_{T3}(47) = -0.68, p = 0.00$) to the post-test ($r_{T2}(66) = -0.78, p = 0.00$; $r_{T3}(50) = -0.68, p = 0.00$). The learners' attention is successfully called towards the falling part of the F0 after finishing the perceptual training since the negative relationship between reaction time and predicted probability of the tone categories is almost as strong as for natives in the learners' post-test.

5. Discussion

Taking all the analysis above together, we would like to speculate the learning progress of L2 learners perceiving Mandarin tone contrasts. The ability of attending to the key acoustic features is an essential step for L2 learners to acquire the ability of categorizing the speech contrasts of the target language. It happens along with the increasing attention to the key acoustic features and the fading out of the other acoustic features they used to use. Although as shown in the results of this study, the fading out of the L2 learners attention on less necessary acoustic features might be a much slower progress than the increasing attention on the key features in the perceptual training condition, the L2 learners' proficiency of accurately discriminating the speech contrasts increases rapidly. Raising the attention to the key acoustic features also has the effect of shifting the perceptual boundary location between the speech contrasts in L2 learners' perceptual space. By relatively extending the perceptual space of one speech contrast while shrinking the other, the L2 learners could gain a more and more native like perceptual space.

On the other hand, L2 learners' ability to generalize the new developed categories to discriminate the speech contrasts in natural speech with rich context and multiple talkers are expected after perceptual training. This ability of generalization relates to the improvement of the robustness of the new developed categorical perception, and it certainly requires much more effort for L2 learners during language learning. We have had taken consideration of L2 learners' generalization ability in designing the two-step perceptual training. The L2 learners in the current experiment were trained by over one hundred natural speech tokens produced by two talkers every day during the second training section. The identification responses of the sixty natural stimuli in the pretest, middle-test and post-test will be analyzed in order to guide us to a better understanding of the learning progress of the robustness of L2 learners' categorical perception.

6. Conclusions

In the present study, we investigated Japanese-native-speaking L2 learners' perception progress of Mandarin Tone 2 and 3 contrasts during a perceptual training experiment. By successfully drawing attention to the key acoustic features, the F0 height differences and length of the falling part of the concave F0 shape, L2 learners gain a significant improvement on shifting the boundary location to a more native like F0 range. On the contrary, the robustness of categorization is not well improved and less necessary acoustic features still affect the perception even after the perceptual training for L2 learners. Future studies are needed to observe L2 learners' perceptual improvement in long-term education programs since we believe that the development of categorical perception requires not only intensive practice but also necessary time for psychological processing.

7. Acknowledgment

The first author would like to acknowledge the help provided by Nicolas Loerbroks for revising the English as well as giving the technical support.

8. References

[1] T. Zou, J. Zhang, and W. Cao, "A comparative study of perception of tone 2 and tone 3 in mandarin by native speakers and japanese learners," in *Chinese Spoken Language Processing (ISCSLP), 2012 8th International Symposium on*. IEEE, 2012, pp. 431–435.

[2] C. B. Moore and A. Jongman, "Speaker normalization in the perception of mandarin chinese tones," *The Journal of the Acoustical Society of America*, vol. 102, no. 3, pp. 1864–1877, 1997.

[3] X. S. Shen and M. Lin, "A perceptual study of mandarin tones 2 and 3," *Language and speech*, vol. 34, no. 2, pp. 145–156, 1991.

[4] J. Gandour, "Tone dissimilarity judgments by chinese listeners," *Journal of Chinese Linguistics*, pp. 235–261, 1984.

[5] B. Chandrasekaran, P. D. Sampath, and P. C. Wong, "Individual variability in cue-weighting and lexical tone learning," *The Journal of the Acoustical Society of America*, vol. 128, no. 1, pp. 456–465, 2010.

[6] N. Loerbroks, Y. Sun, Y. Sagisaka, and J. Zhang, "Visualization of mandarin chinese tone production of japanese l2 learners for evaluation," *Language Teaching, Learning and Technology*, pp. 1–5, 2016.

[7] D. H. Whalen and Y. Xu, "Information for mandarin tones in the amplitude contour and in brief segments," *Phonetica*, vol. 49, no. 1, pp. 25–47, 1992.

[8] Y. Wang, M. M. Spence, A. Jongman, and J. A. Sereno, "Training american listeners to perceive mandarin tones," *The Journal of the Acoustical Society of America*, vol. 106, no. 6, pp. 3649–3658, 1999.

[9] 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, 2003.

[10] J. Zhang, X. Wang, Y. Sun, M. Nishida, T. Zou, and S. Yamamoto, "Improve japanese c2l learners' capability to distinguish chinese tone 2 and tone 3 through perceptual training," in *Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference*. IEEE, 2013, pp. 1–6.

[11] D. G. Jamieson and D. E. Morosan, "Training new, nonnative speech contrasts: A comparison of the prototype and perceptual fading techniques." *Canadian Journal of Psychology/Revue canadienne de psychologie*, vol. 43, no. 1, p. 88, 1989.

[12] J. S. Logan, S. E. Lively, and D. B. Pisoni, "Training japanese listeners to identify english/r/and/l: A first report," *The Journal of the Acoustical Society of America*, vol. 89, no. 2, pp. 874–886, 1991.

[13] J. S. Logan, S. E. Lively, and D. B. Pisoni, "Training listeners to perceive novel phonetic categories: How do we know what is learned?" *The Journal of the Acoustical Society of America*, vol. 94, no. 2, pp. 1148–1151, 1993.

[14] W. Schneider, A. Eschman, and A. Zuccolotto, *E-Prime: User's guide*. Psychology Software Incorporated, 2002.

[15] G. S. Morrison, "Logistic regression modelling for first and second language perception data," *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, vol. 282, p. 219, 2007.

[16] G. S. Morrison and M. V. Kondaurova, "Analysis of categorical response data: Use logistic regression rather than endpoint-difference scores or discriminant analysis," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2159–2162, 2009.

[17] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017. [Online]. Available: https://www.R-project.org/

[18] R. M. Nosofsky, "Similarity scaling and cognitive process models," *Annual review of Psychology*, vol. 43, no. 1, pp. 25–53, 1992.

[19] Y. Sun, J. Zhang, and Y. Sagisaka, "Measuring reaction time of chinese tone identification for finer. evaluation of l2 learner's proficiency," in *Proceedings of the 18th Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation*. IEEE, 2015, pp. 300–304.