5-2019

The Perception of Mandarin Tones in "Bubble" Noise by Native and L2 Listeners

Mengxuan Zhao

The Graduate Center, City University of New York

Recommended Citation

This Thesis is brought to you by CUNY Academic Works. It has been accepted for inclusion in All Dissertations, Theses, and Capstone Projects by an authorized administrator of CUNY Academic Works. For more information, please contact deposit@gc.cuny.edu.
THE PERCEPTION OF MANDARIN TONES IN “BUBBLE” NOISE BY NATIVE AND L2 LISTENERS

by

MENGXUAN ZHAO

A master’s thesis submitted to the Graduate Faculty in Linguistics in partial fulfillment of the requirements for the degree of Master of Arts, The City University of New York

2019
The Perception of Mandarin Tones in “Bubble” Noise by Native and L2 Listeners

by

Mengxuan Zhao

This manuscript has been read and accepted for the Graduate Faculty in Linguistics in satisfaction of the thesis requirement for the degree of Master of Arts.

Date

Michael I. Mandel
Thesis Advisor

Gita Martohardjono
Executive Officer

The City University of New York
ABSTRACT

THE PERCEPTION OF MANDARIN TONES IN “BUBBLE” NOISE BY NATIVE AND L2 LISTENERS

by

MENGXUAN ZHAO

Adviser: Michael I. Mandel

Previous studies have revealed the complexity of Mandarin Tones. For example, similarities in the pitch contours of tones 2 and 3 and tones 3 and 4 cause confusion for listeners. The realization of a tone’s contour is highly dependent on its context, especially the preceding pitch. This is known as the coarticulation effect. Researchers have demonstrated the robustness of tone perception by both native and non-native listeners, even with incomplete acoustic information or in noisy environment. However, non-native listeners were observed to behave differently from native listeners in their use of contextual information. For example, the disagreement between the end of a preceding pitch and the beginning of a target pitch degraded L2 listeners’ identification performance, but not native listeners’. In this thesis, we use the “bubble” noise techniques (Mandel et al., 2016) to explore the difference between native and L2 listeners’ tone perception. We focus on whether the two groups of listeners use different time-frequency cues to identify tones, and especially if they use cues from preceding tones differently. We conducted a tone identification listening test on 5 native listeners and 5 L2 listeners. Stimuli consisted of 6 meaningful sentences with the target syllable at the end. The target syllable used one of tones 2, 3 and 4 and was preceded by either tone 1 (high) or 4 (low). These sentences were mixed with “bubble” noise. Our results showed some difference in the use of preceding context between the two groups. Statistical analysis of the difficulty
level of the task supported that this difference is prominent in harder listening tasks and not in easier tasks.
ACKNOWLEDGMENTS

First, I would like to express my gratitude to my advisor Professor Michael I. Mandel. Without his insightful guidance and persistent help, I would not have accomplished this thesis. His advice on stimuli design and experiment result analysis resulted in the core of this paper. It was such an invaluable experience to work with him. I am proud of what I have learned and gained under his advisement. I would also like to thank Professor Martin Chodorow who advised on the statistical analysis and Professor Kyle Gorman who was always willing to answer my technical and statistical questions.

I am thankful to the Second Language Acquisition Lab. Participating in one of its research projects gained me precious experience in working on experimental research. Additionally, my colleagues at the lab, as well as our lovely assistant program officer Nishi Bissoondial, generously helped me with finding places to run my own experiments.

Many thanks to those who participated in my pilot experiments. Some of them participated multiple times. They are Xiaomeng Ma, Ben Macaulay, Armando Tapia, Anthony Vicario, Lara Novic, Ariel Silkett, Chen Zhou and Rachel Rakov from the Linguistics department, and Jiaqi Yang, Shao Liu and Yimeng Guo from the Computer Science department. Their voluntary participation helped me improve my experiment designs. In particular, Ben gave me some crucial suggestions on the stimuli design. Thanks to Dr. Chun-Yi Peng at Borough of Manhattan Community College and the Chinese Division of Hunter College for their help in recruiting experiment participants.

I am grateful for my friends at the Linguistics department and my family in Beijing. They unconditionally and emotionally supported me throughout my whole journey of completing this master’s degree.
# Contents

Contents vii

List of Tables ix

List of Figures x

1 Introduction 1

1.1 Contour shape and tone confusion ........................................ 2
1.2 Context and coarticulation .................................................. 4
1.3 Acoustic cues and robustness of tone perception .......................... 5
1.4 Nativeness and tone perception ............................................. 7
1.5 Summary ........................................................................... 9

2 Experiment 10

2.1 Stimuli ............................................................................... 11
2.2 Bubble noise methodology .................................................. 13
2.3 Hypotheses ......................................................................... 15
2.4 Participants ......................................................................... 15
2.5 Procedure .......................................................................... 16

3 Results 18

3.1 Significant areas ............................................................... 18
3.1.1 Importance maps ............................................................ 18
3.1.2 Comparison by nativeness ............................................... 20
3.2 Final BPS .......................................................................... 21
List of Tables

1. Tone representations .................................................. 2
2. Stimuli ................................................................. 11
3. Final BPS levels ......................................................... 22
4. Native listeners results ................................................. 32
5. L2 listeners results ..................................................... 34
# List of Figures

1. Five-scale representation ........................................ 1
2. Carry-over effects ................................................. 12
3. Example visualizations of correlation analysis ................ 14
4. Example MATLAB interface ........................................ 16
5. Importance maps: native ........................................... 19
6. Importance maps: L2 ................................................. 19
7. Significant differences between correlations .................. 21
8. Analysis: final BPS by context .................................... 23
9. Analysis: final BPS by tone ........................................ 23
1 Introduction

Mandarin Chinese is a tonal language and it uses tones as suprasegmental features, in addition to segmental features such as consonants and vowels, to construct syllables and represent meanings. Tones carry no less functional load than vowels (Surendran and Levow, 2004). Every stressed syllable has a tone spread over the voiced part of the syllable (Chao, 1968). There are four lexical tones in Mandarin. A simple way to mark the four lexical tones is the five-scale representation, introduced by Chao (1968), which is widely used for academic research and pedagogical purposes. In this system, 1 represents the lowest pitch and 5 the highest. The four tones can then be represented based on their general pitch contour as illustrated in Figure 1 and shown in Table 1 (Chao, 1968; Howie and Howie, 1976). There is also a neutral tone, whose pitch is highly dependent on the preceding context (Duanmu, 2007), however the neutral tone is out of this thesis’s scope of discussion.

![Figure 1: Five-scale representation by Chao (1968)]

In this thesis, we also mark tones using pinyin notation (Ministry of Education of the People’s Republic of China, 1958). Tones are marked on the main vowel of the syllable as: ã (tone 1), ã (tone 2), ã (tone 3), and ã (tone 4).
Table 1: Pinyin notation, five-scale representation, and description of the four Mandarin tones.

<table>
<thead>
<tr>
<th>Tone</th>
<th>Pinyin</th>
<th>Five-scale</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>å</td>
<td>55</td>
<td>high-level</td>
</tr>
<tr>
<td>2</td>
<td>á</td>
<td>35</td>
<td>rising</td>
</tr>
<tr>
<td>3</td>
<td>å</td>
<td>214</td>
<td>low-dipping</td>
</tr>
<tr>
<td>4</td>
<td>à</td>
<td>51</td>
<td>falling</td>
</tr>
</tbody>
</table>

1.1 Contour shape and tone confusion

Tone 3 has several variations when produced in different contexts. When appearing at a non-final position, or sometimes in casual speech, the rising part of tone 3 is missing and only the falling onset remains (Chao, 1968; Shih, 1988; Duanmu, 2007; Yang, 2010). This is known as the “half tone 3” sandhi.

The notations of tone 2 (35) and tone 4 (51) indicate that the contour rises or falls directly as a straight line, but this is not precisely true. Acoustic studies show that tone 2 usually has a slight falling slope at its onset, resulting in a concave shape similar to tone 3 (Shen and Lin, 1991; Xu, 1997; Yang, 2010). Similarly, tone 4 has a small initial rise (Gårding et al., 1986; Xu, 1997). Because of the similarities in contour shapes, tones 2, 3 and 4 tend to cause confusions in tone perception (Gårding et al., 1986; Blicher et al., 1990; Shen and Lin, 1991; Gottfried and Suiter, 1997; Jongman and Moore, 2000; Liu and Samuel, 2004; Lee et al., 2008, 2010a; Hao, 2018a). Specifically, tones 2 and 3 easily cause confusion in identification tasks because both have a concave pitch contour shape. In most previous studies, they are the most confusing pair (Gottfried and Suiter, 1997; Liu and Samuel, 2004; Lai and Zhang, 2008; Lee et al., 2008, 2010a; Hao, 2018b).

Blicher et al. (1990) conducted two experiments to explore how duration influences the discrimination of tones 2 and 3. In the first experiment, the stimuli were contours of a typical tone 2, a typical tone 3 and contours in between the two. Participants had to choose one
of the two tones based on what they heard. Results showed that proportionally lengthening F0 contours that are intermediate between tones 2 and 3 led them sound more like tone 3. This was both true to native Mandarin listeners and native English listeners. In the second experiment, participants judged whether one syllable in the syllable pairs were longer than the other. It was observed that tone 3 was perceived as longer than tone 2 even when they were equal in duration, and actual lengthening reinforces this apparent-length cue for tone 3.

Apart from syllable duration, Shen and Lin (1991) and Jongman and Moore (2000) further explored the role of turning point, F0 difference and speaking rate in tone 2/3 discrimination tasks. Shen and Lin (1991) manipulated the turning point of two F0 continua, and had native Mandarin listeners judge if they heard tone 2 or tone 3. Results showed that there was a correlation between the timing and the F0 difference between the onset and the turning point, and listeners chose tone 2 or 3 based on the interaction of both factors. In identification tasks, errors were mostly due to the ambiguity derived from the correlation. Jongman and Moore (2000) tested on rate normalization and speaker normalization in tone perception by native Mandarin listeners and American English monolingual listeners. It was found that Mandarin listeners normalized for both speaker F0 range and speaking rate when precursors and stimuli varied in the same acoustic dimension. However, for English listeners, normalization happened as a result of both F0 and turning point change, which could be regarded as pure acoustic discrimination.

Since both tone 3 and tone 4 have a falling contour, Gårding et al. (1986) studied factors that influence the discrimination of the two. With pairs of sentences that were different only by one syllable, researchers manipulated the tone 3/4 contour by changing the position and the pitch height of the peak, the steepness of the falling contour, and adding creaky voice. Results showed that manipulation of F0 changed tone 4 to tone 3, while creak was not necessary in the perception of tone 3.
### 1.2 Context and coarticulation

Based on the relative pitch height of onsets and offsets, the contour patterns of the four tones are regarded as “high-high”, “low-high”, “low-low” and “high-low” (Shih, 1988). Two similar adjacent pitch values construct a “compatible context” and two different adjacent pitch values construct a “conflicting context”. The realization of tone contours in real speech is influenced by the contextual condition. Sometimes, the pitch contour can be changed by surrounding tones. This is known as the tone coarticulation effect. There are two types of coarticulation effects. The carry-over effect is the influence of a tone on the following tone, and the anticipatory effect is the influence of a tone on the preceding tone.

An earlier study by Shen (1990) concluded that both carry-over and anticipatory effects symmetrically affect Mandarin tones. The effect of coarticulation was on the entire F0 contour, not only the onset or offset. Coarticulation effects influenced only the two adjacent tones, but did not extend to a wider range of context.

Xu (1993, 1994a,b, 1997) systematically studied the coarticulation of Mandarin tones. Xu argued that the identification of tones in a compatible context was highly robust with or without the context, while identification performance was diminished in conflicting context with the absence of the contextual information. In other words, if a syllable appears in compatible context, it is equally easy to identify its tone by itself or with contextual information. On the other hand, if a syllable is represented in conflicting context, identifying its tone in isolation is much harder than with contextual information. This is to say, listeners take advantage of the context so as to compensate for the coarticulation effects. Xu compared the carry-over effect and the anticipatory effect, and found that the carry-over effect is assimilatory while another effect is dissimilatory. For example, in the carry-over effect, when the preceding tone ends at a high pitch, the onset of the following syllable rises and vice versa. However, in the anticipatory effect, if the following syllable ends at a high
pitch, the preceding syllable tends to become lower in F0, and vice versa. Xu argued against Shen (1990) and asserted that the carry-over effect is asymmetrically more noticeable than anticipatory effect. Xu (1997) used balanced bi-tonal sequences and established a baseline for contextual tone coarticulation in Mandarin.

Huang and Holt (2009) claimed that tone perception is highly context-dependent. They used precursors of different F0s before a tone 1/2 contrast, and observed that both speech and non-speech precursors affected tone perception. With a higher-frequency precursor, the following tone is more likely to be perceived as a lower-frequency tone. Lee et al. (2010a) found that native Mandarin listeners have normal recognition performance in different contextual conditions, while L2 listeners’ performance decreased when a syllable was originally articulated in a conflicting context and presented by itself. This indicated that native listeners are better able to compensate for coarticulation effects than L2 listeners.

1.3 Acoustic cues and robustness of tone perception

Tone perception is primarily based on continuous acoustic properties, especially pitch information, although secondary auditory cues such as amplitude and duration are also important for tone identification (Tseng et al., 1986; Norman, 1988; Whalen and Xu, 1992; Liu and Samuel, 2004; Yang, 2010). Other cues include the timing and the difference in F0 from onset to turning point (Blicher et al., 1990; Shen and Lin, 1991; Jongman and Moore, 2000), talker-to-listener distance (Shih and Lu, 2015) etc.

Tseng et al. (1986) conducted experiments on tone identification by native Mandarin listeners using a synthesized carrier vowel /i/ with pre-set duration and formant frequencies. F0 was specified at 11 points of the vowel, and the task was to identify if it is tone 1 or tone 2. They found that tone perception requires both F0 contour and F0 height, and the contribution of one increases when the other is ambiguous. Huang and Johnson (2010) and
Guion and Pederson (2007) pointed out that language experience influences how listeners attend to these two acoustic dimensions. They found that native Mandarin listeners were more sensitive to F0 contour while English listeners, on the other hand, attended more to F0 height.

The duration of the four tones are different in nature (Nordenhake and Svantesson, 1983), which makes duration effective in assisting tone identification. Liu and Samuel (2004) defended that tone duration played a significant role when F0 information was missing. Their participants’ performance on tones 1, 3, and 4 were significantly better than chance when they were tested on whispered stimuli, where no F0 information existed.

Whalen and Xu (1992) adopted signal-correlated noise stimuli to test how amplitude contour and duration influence tone perception when F0 information is unavailable. Results showed that tones 2, 3, and 4 could be perceived by amplitude contour exclusively, indicating that amplitude contours can be used as tone identification cues. Additionally, F0 was found to be positively correlated with amplitude, meaning that not every portion of the F0 equally and unambiguously reveals the tone.

Because listeners can use various acoustic cues, tone perception of Mandarin remains robust even in adverse conditions due to partial absence of a syllable or the presence of noise. Liu and Samuel (2004) observed that native listeners’ identification accuracies were far above chance when partial F0 information was neutralized, and were still above chance when F0 information was completely lost except for tone 1. Gottfried and Suiter (1997) and Lee et al. (2008, 2009, 2010a,b, 2013) used intact, center-only, silent-center, and initial-only syllables, where they removed six pitch periods from the beginning and eight pitch periods from the end, or remain the beginning and the end exclusively, or remain only the beginning. When the stimuli were presented without any noise, both native and L2 listeners performed above chance in all four conditions. Dees et al. (2007) also showed the robustness of tone identification in white noise and multi-talker babble noise.
1.4 Nativeness and tone perception

Studies have shown differences between native and non-native listeners in tone perception. With even no background in tonal language at all, listeners are still able to distinguish Mandarin tonal contrasts with good accuracy (Guion and Pederson, 2007; Dees et al., 2007; Huang and Johnson, 2010; Hao, 2018b). In this thesis, we are more interested in the difference between beginning level language learners (L2 listeners) who have limited experience in Mandarin and native Mandarin listeners. Previous studies have found that native and L2 listeners assign different weights to different acoustic cues (Guion and Pederson, 2007), are influenced differently by context and other external factors such as noise (Gottfried and Suiter, 1997; Lee et al., 2008, 2009, 2010a,b, 2013), and show different error patterns (Gottfried and Suiter, 1997; Lee et al., 2008, 2009, 2010a,b, 2013; Hao, 2018b).

Guion and Pederson (2007) compared the tone perception between native Mandarin speakers, native English monolinguals, native English advanced-level Mandarin learners who begun studying Mandarin as an adult, and native Japanese speakers. Their results showed that native Mandarin listeners relied on both F0 height and contour shape, while the monolingual English group and the Japanese group did not use much contour shape information. On the other hand, language learning experience changed the weighting of acoustic cues. The Mandarin learner group had learned to use both acoustic cues. Based on their findings, we can assume that different from advanced-level learners, beginning-level learners are experiencing a transition in perceptual strategy and their tone perception shows some interesting complexity.

Hao (2018b) tested Mandarin vowel and tone discrimination on English listeners with no, little, and advanced experience in Mandarin. It was corroborated that language learning experience contributed to the improvement in tonal contrast identification accuracy. However, the contribution was not as effective for the tone 2/3 contrast as for other pairs. In
other words, experienced L2 learners achieve native-like accuracy for only some of the tone contrast pairs. This indicates that apart from the familiarity of pitch and contour patterns of the tones, there are other factors causing the difference between native and L2 listeners’ tone perception.

Gottfried and Suiter (1997) conducted several listening tests on both native and L2 listeners. They used intact, center-only, silent-center, and initial-only syllables with or without the following syllable /tsi/. In both conditions, native listeners achieved high accuracy for all syllables except initial-only, while L2 performance decreased as a function of information retained in the syllables. This discrepancy was salient for silent-center syllables. Native listeners were able to make judgement based on the combination of onset and offset information, however L2 listeners were not as adept in making use of this information. Another difference between the two groups was their error patterns. Native listeners mainly made confusions between tones 2 and 3, both having lower onset, or tones 1 and 4, both having higher onset. L2 listeners’ error patterns varied. For example, tone 3 and 4 confusion which was rare among native listeners, occurred comparatively frequently in the L2 group.

Lee et al. (2008, 2009, 2010a,b, 2013) conducted a series of studies using the four types of stimuli in Gottfried and Suiter (1997), focusing on the differences in Mandarin tone perception between native and L2 listeners. Considering the asymmetry of the coarticulation effect, Lee et al. (2008, 2010a) instead of following context, used two preceding contexts, one with a high pitch offset and the other with a low pitch offset. Their results showed that the existence of context facilitated both groups. While native listeners performed consistently well in both contexts, L2 listeners performed better in the compatible context and were disrupted by the conflicting context where the preceding tone and the target tone were not continuous in F0. This is to say, L2 listeners were not as sensitive to contextual tonal variability as native listeners were, relying primarily on F0 information of the target syllable itself in tone identification. Results from Lee et al. (2009) also supported this conclusion.
Lee et al. (2009, 2010b, 2013) discussed the effect of multiple speaker and noise on tone perception. Lee et al. (2010b, 2013) argued that listeners with higher Mandarin proficiency are more negatively affected by noise in tone perception, but native listeners still achieved higher accuracy in all noise levels. However, this might be due to that less proficient listeners’ performance was initially low, thus the extend to which the performance could be reduced was limited.

1.5 Summary

Previous studies have discussed the acoustic and phonetic features of Mandarin tones and factors related to tone perception, as well as compared native and non-native listeners in tone perception. They found that native Mandarin speakers and people from non-tonal language background who speak Mandarin as a foreign language perceive tone differently. These differences are manifested as different error patterns and different sensitivity to context and missing information. In particular, we are interested in the facts that native and L2 listeners are differently influenced by additive noise and have different capability in compensating for coarticulation effects. In this thesis, we will continue exploring the influences and interaction among tone, context and nativeness in Mandarin tone perception.
2 Experiment

According to previous studies, native and L2 listeners perform differently in tone identification tasks. Specifically, their abilities vary in compensating for the coarticulation effect. In this thesis, we explore these differences using “bubble” noise techniques. In short, we expect to answer the following questions:

1. Which acoustic cues are used by native listeners when they identify Mandarin tones in a noisy environment?

2. Which acoustic cues are used by L2 listeners when they identify Mandarin tones in a noisy environment?

3. Do listeners with different language backgrounds use different acoustic cues for tone perception?

4. Do listeners use different cues for different tones and in different contexts?

The goal of this study is to locate in time-frequency spectrograms the acoustic cues that listeners use to correctly identify tones, and to compare the differences between these cues by tone, context, and most importantly, nativeness. The tones we study are not produced in isolation. Instead, they are produced in carrier phrases with specific preceding contexts. In particular, we intend to explore how the carry-over effect influences tone perception. In other words, how do listeners make use of preceding context in tone identification and do the two groups of listeners use preceding context differently? With two types of context (high and low) and three target tones (tones 2, 3, and 4), the stimuli provide six combinations. Of the six combinations, three are compatible contexts and three are conflicting contexts. In Section 2.1, we will introduce the motivation for the design of the stimulus sentences.

To investigate the above research questions, we use the bubble noise framework of Mandel et al. (2016). This framework generates noise with bubbles that reveal glimpses of the original
speech utterances at random time-frequency positions. Based on human judgments, the system computes which positions are crucial to the correct identification of the speech. This mechanism allows us to find the time-frequency cues that different listeners use to identify the tones, and to compare the different use of cues under different conditions. In Section 2.2, we will introduce more details about the bubble noise methodology.

In the rest of Section 2, we will introduce our hypotheses, present information about the participants, and describe the experimental procedures.

### 2.1 Stimuli

We used six stimulus sentences in our experiment (see Table 2). All six sentences start with the identical phrase “tā jiào”, meaning “his/her name is”. The last syllable “wei” is the target. It is one of tones 2, 3, and 4. The third syllable provides two types of contexts, high offset (“zhú”) or low offset (“zhù”). All the six “zhu wei” tone combinations are valid names in Mandarin.

Our stimuli are adapted from the stimuli of Lee et al. (2008, 2010a) who used a pair of carrier phrases “qǐng shuō _” and “qǐng kàn _”, providing two different contextual conditions for the target tone. One problem is that their target syllable and the context syllable are at different levels of the syntactic hierarchy, and there is possibly an intonational pause between them. Thus we used names to avoid the potential intonational pause. Another reason for

<table>
<thead>
<tr>
<th>Target</th>
<th>Preceding context</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High (Tone 1)</td>
</tr>
<tr>
<td>Tone 2</td>
<td>tā jiào zhù wéi</td>
</tr>
<tr>
<td>Tone 3</td>
<td>tā jiào zhù wéi</td>
</tr>
<tr>
<td>Tone 4</td>
<td>tā jiào zhú wéi</td>
</tr>
</tbody>
</table>

Table 2: List of sentences used in Mandarin tone perception experiment showing the tone of the target word “wéi”, “wēi”, or “wèi” in the context of the tone of the preceding word “zhū” or “zhù”. The sentences translate to “His/Her name is zhù wéi” for example.
using names is to minimize the influence of word familiarity between native and L2 listeners. We assume that listeners from both groups find the names unfamiliar yet still meaningful.

Figure 2 shows the F0 contours of the six stimulus sentences. Carry-over effects can be observed from the illustration. The part of the contours corresponding to the last syllable in light gray are higher than those in dark gray in general, because they are influenced by the higher offset of the preceding tone 1. The pitch contours were extracted using Python Swipe package. Swipe is an open source software for speech pitch extraction based on Camacho (2007). Code is available: https://github.com/kylebgorman/swipe

Additionally, we used the “half tone 3” in the two tone 3 sentences. As is mentioned in Section 1, when appearing at a non-final position, or sometimes in casual speech, the rising part of tone 3 is missing (Chao, 1968; Shih, 1988; Duanmu, 2007; Yang, 2010). Considering that half tone 3 appears at non-end positions and in casual speech when people speak faster, it is reasonable to assume that half tone 3 in general appears no less or even more frequently than full tone 3 in speech. In addition, we used the semi-vowel /w/ at the beginning of the target syllables to better illustrate the coarticulation effects, because a semi-vowel retains more F0 information than other consonants such as plosives or fricatives.

All sentences were recorded in their entirety to provide natural speech. Each sentence was recorded several times and one of each was selected so as to most closely match the selected version of the other sentences in timing and loudness. Listeners were only asked to
identify the target syllable.

2.2 Bubble noise methodology

Mandel et al. (2016) introduced a framework to identify the time-frequency locations of cues that listeners use to reliably identify continuous utterances. By using the same speech utterance mixed with different bubble noise instances, the framework computes the importance to the intelligibility of each point on the time-frequency spectrogram.

Following Mandel et al. (2016), in our experiment, speech utterances of short sentences in Mandarin Chinese are each combined with multiple instances of bubble noise. The noise is derived from speech-shaped Gaussian noise with a signal-noise-ratio of -25 dB that makes the speech completely unintelligible. These bubble-shaped glimpses are attenuated so that at these bubble locations the original speech is revealed. The bubbles are produced at random locations in time and in ERB<sub>N</sub>-scale frequency (Glasberg and Moore, 1990) with a maximum suppression of 80 dB. The mathematical representation of the attenuation applied to the bubble noise is

\[
B(f,t) = \sum_{i=1}^{I} \exp \left\{ -\frac{(t - t_i)^2}{\sigma_t^2} - \frac{(E(f) - E(f_i))^2}{\sigma_f^2} \right\} \tag{1}
\]

\[
M(f,t) = \min \left( 1, \frac{10^{-80/20}}{B(f,t)} \right) \tag{2}
\]

where \(E(f) = 21.4 \log_{10}(0.00437f + 1)\) converts frequencies in Hz to ERB<sub>N</sub>, and \(\{(f_i, t_i)\}_{i=1}^{I}\) are the randomly selected centers of the \(I\) bubbles. The scale parameters \(\sigma_t\) and \(\sigma_f\) were set such that the bubbles were fixed in size to have a half-amplitude “width” of 90 ms at their widest and a half-amplitude “height” of 1 ERB<sub>N</sub> at their tallest, the smallest values that would avoid introducing audible artifacts. For the full 80-dB dynamic range, this corresponds
Figure 3: The original spectrogram, the correlation plot, and the significant correlations of stimuli sentence “tā jiào zhù wèi” identified by native listeners. After the significance test and the false detection rate control, it is shown that the significant positive correlations are mostly located from 1200 ms to 1600 ms at the low frequency area.

to 350 ms wide at their widest and 7 ERB$_N$ at their highest.

In order to analyze the experimental results, the framework computes the correlation between the audibility at each individual time-frequency point and the correct identification of the speech using the point-biserial correlation (Calandruccio and Doherty, 2007). This is the correlation between a dichotomous variable (correct or incorrect identification) with a continuous variable (audibility at a time-frequency point). The significance, denoted $p(f, t)$ is also tested using a one-way analysis of variance with two levels. The significant correlations are considered important to the intelligibility of that word (see Figure 3).

To control the false detection rate due to considerable amount of parallel statistical tests, the method of Benjamini and Hochberg (1995) and Benjamini et al. (2001) are applied. In particular, when performance many statistical tests in parallel, as the bubble noise system does, this method corrects the results so that only a fixed proportion of identified significant results may be spurious. The system uses a proportion of 5%, meaning that if 1000 important time-frequency points are identified out of 100,000, then at most 50 of those identifications will be spurious. The original method of Benjamini and Hochberg (1995) is appropriate when the tests are positively correlated or uncorrelated and the method of Benjamini et al. (2001) is appropriate if the tests can be negatively correlated.

Experiments can be carried out using an adaptive or non-adaptive procedure. The non-
adaptive procedure keeps the number of bubble-per-second (BPS) constant. The adaptive procedure uses the weighted up-down procedure (Kaernbach, 1991) to adjust the number of BPS to achieve approximately 60% accuracy. In this procedure, each time the listener makes a correct choice, the task becomes slightly more difficult by decreasing its BPS, while each time the listener makes an incorrect choice, the task becomes slightly easier by increasing its BPS.

Using this framework, Mandel et al. (2016) studied the perception of several English consonants by native English speakers. Later Choi (2018) adopted the methodology and compared the perception of Korean fricatives and affricates by native and non-native Korean listeners.

2.3 Hypotheses

We have several hypotheses for this experiment. First, according to Gottfried and Suiter (1997) and Lee et al. (2008, 2009, 2010a,b, 2013), native listeners are better able to use contextual and coarticulation information in adverse environments. We expect to see this difference in the importance maps generated by the bubble noise framework. Second, we expect native listeners to achieve higher accuracy in tone identification tasks, with or without added noise. Considering that the adaptive procedure adjusts the difficulty to achieve the target accuracy, we expect to find that the two groups end at different difficulty levels, represented as different BPS levels. In addition, we also want to know if different tones show different patterns in importance maps.

2.4 Participants

Ten subjects participated in total, five native listeners and five L2 listeners. The native listeners were graduate students from the Graduate Center, City University of New York.
Figure 4: Example MATLAB interface. Since all participants were familiar with mandarin tones, we used numbers instead of pinyin tone markers to indicate each tone. Participants needed to type “1” for tone 2, “2” for tone 3, “3” for tone 4, or “4” for “don’t know”. After making the choice, they pressed “enter”, and the next sentence would be played and the next question would show up on the screen.

and New York University, with an average age of 26.4. All were native speakers of Mandarin and did not have a noticeable dialect accent. Their second language was English. L2 listeners were undergraduate students from Borough of Manhattan Community College and Hunter College, with an average age of 19. They had studied Mandarin for less than 2 years. One was monolingual English speaker. Four were English-Spanish bilingual speakers. None spoke any other tonal language apart from Mandarin. Most of them were compensated for participating the experiment, and the rest participated voluntarily. None of the participants reported any hearing problem.

2.5 Procedure

The experiment took place in a quiet room. Participants were seated in front of a laptop with a MATLAB interface. The stimuli were presented diotically. Participants adjusted the volume to a comfortable level. While the experiment was on going, the participants could take small breaks whenever needed.

The experiments were composed of three sessions of forced choice tasks. In all of the three sessions, they heard each of the six stimulus sentences in random order, and were asked to choose the tone of syllable “wei” in the MATLAB interface. Each option, the three tones and “don’t know”, corresponded to a numbered key on the keyboard. Even though they could replay the sound as many times as they would like, they were encouraged to make
judgements based on their intuition and not to spend unnecessary time on each question. The first two were pilot sessions that helped the participants get familiar with the system. Pilot session I was conducted only on the L2 group. In this session, sentences were played with no noise and with feedback, each six times in random order. Each time they made a selection, they were told if they chose the correct answer.

Pilot session II was conducted on all participants. Sentences were played in non-adaptive mode with feedback. Listeners heard sentences in bubble noise at a BPS of 15, each for six times in random order. Feedback were given after each choice.

In the last session, sentences were played in adaptive mode and without any feedback, with a initial BPS of 15. 200 different bubble noise instances were generated randomly for each sentence. Over all, participants heard 1,200 speech-noise mixtures, and each had its own random bubble noise pattern. These sentences were presented in random order blocked in groups of 24. The whole procedure took approximately 1.5 to 2 hours to complete.
3 Results

3.1 Significant areas

3.1.1 Importance maps

Combining the results from all five participants in each group, an importance map was generated for each sentence, shown in Figures 5 and 6. Brighter regions show significant correlations between audibility and correct identification.

As is shown in the plots of native listeners’ results, for tones 2 and 3, the highlighted area is mainly composed of a lower frequency part under 1 kHz, and a higher frequency part above 1 kHz. It can be seen from native tone 3 plots that the high frequency area goes from 1 kHz to approximately 4 kHz. On the other hand, tone 2 shows smaller importance area of higher frequency. The higher frequency highlight area mainly covers the target syllable, while the lower frequency area extends to the preceding syllable to some extent. As to tone 4, the significant area extends from the offset of the preceding syllable to the offset of the target syllable and it reaches approximately to 2.5 kHz at the onset of the target syllable.

The L2 group shows similar highlighted areas of tone 4 sentences. However, the tone 3 plots of L2 results barely show a conspicuous significant area. Both of tone 2 plots show some weak highlight in high frequency area, while in low frequency area under 1 kHz the highlight is clearer.

The importance maps indicate that both groups use similar acoustic cues in the identification of tone 4. Discrepancies mainly lie in tones 2 and 3. Even though both groups show highlighted areas at approximately the same locations in tone 2 plots, the areas extend wider in native results. The tone 2 highlighted areas of L2 listeners do not contain as much preceding context as the native listeners’ do. The lack of highlighted area in L2 tone 3 plots is possibly due to either this group has difficulty identifying the tone, or each participant
has their own significant area that does not overlap with each others.

Figure 5: Importance maps of native listeners. This is generated by combining the results from all five native listeners.

Figure 6: Importance maps of L2 listeners. This is generated by combining the results from all five L2 listeners.
3.1.2 Comparison by nativeness

To better compare the difference between native and L2 listeners, we adopted the method from Wuensch (2019) and Fisher (1921) to compare if the point-biserial correlations are significantly different between native and L2 subjects. The correlations at each time-frequency point are normalized into z-scores using the Fisher z-transform,

\[ z = 0.5 \log_e \left( \frac{1 + r}{1 - r} \right). \] (3)

The two z-score matrices are compared and result in a new z-score matrix. The one-sided p-value is calculated out of this z-score and corrected by false discovery rate (Benjamini and Hochberg, 1995; Benjamini et al., 2001)

\[ Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}. \] (4)

The results are shown in Figure 7. The tone 4 plots show no significant difference as expected. Tone 2 with high offset context shows a minor significant difference at the onset of the target syllable and approximately below 1 kHz. Tone 2 with low offset context shows a significant difference extending from the onset of the preceding syllable to the middle of the target syllable and below 1 kHz. Tone 3 with high context mainly shows two significant areas, one below 1 kHz and extending from the middle of the preceding syllable to the offset of the target syllable, the other from 1 kHz to nearly 3 kHz, covering mainly the target syllable. Tone 3 with low context also shows a lower frequency area and a higher frequency area, only both are smaller and do not extend to the preceding syllable.

The significant difference between correlations displays how native and L2 listeners weight the cues differently. A significant difference means either that native listeners used the information of a particular area in tone identification while L2 listeners did not, or that both...
used the information of the particular area, yet the correlation was stronger with native listeners than L2 listeners.

3.2 Final BPS

In order to analyze if different contexts and different tones cause different degrees of task difficulty between the two groups, we ran a statistic analysis on the final BPS data (see Table 3). In adaptive mode, the BPS varies in accordance with the task difficulty. In general, a larger BPS means that stimulus is more difficult to recognize and a smaller BPS means it is easier.

A 2 (context: high offset, low offset) × 3 (tone: tone 2, tone 3, tone 4) × 2 (nativeness: native, L2) mixed-design analysis of variance was run on the natural log of the final BPS, with N=5 participants in each nativeness group. Result showed a significant effect in tone 2 (M=2.690, SD=0.359), tone 3 (M=3.114, SD=0.695) and tone 4 (M=1.932, SD=0.555),
Table 3: Final BPS levels

<table>
<thead>
<tr>
<th>Participant</th>
<th>Nativeness</th>
<th>zhū wéi</th>
<th>zhū wéi</th>
<th>zhū wéi</th>
<th>zhū wéi</th>
<th>zhū wéi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>L2</td>
<td>10.712</td>
<td>27.714</td>
<td>221.670</td>
<td>39.582</td>
<td>8.4467</td>
</tr>
<tr>
<td>5</td>
<td>L2</td>
<td>26.1150</td>
<td>29.410</td>
<td>53.272</td>
<td>53.272</td>
<td>2.7320</td>
</tr>
<tr>
<td>6</td>
<td>native</td>
<td>10.0950</td>
<td>10.927</td>
<td>17.926</td>
<td>16.892</td>
<td>4.5717</td>
</tr>
<tr>
<td>7</td>
<td>native</td>
<td>14.4180</td>
<td>15.300</td>
<td>21.852</td>
<td>18.651</td>
<td>11.3680</td>
</tr>
<tr>
<td>9</td>
<td>native</td>
<td>9.5123</td>
<td>12.305</td>
<td>19.792</td>
<td>15.300</td>
<td>2.8992</td>
</tr>
<tr>
<td>Avg</td>
<td>native</td>
<td>12.10146</td>
<td>13.1086</td>
<td>17.3496</td>
<td>15.4348</td>
<td>8.3745</td>
</tr>
<tr>
<td>Avg</td>
<td>overall</td>
<td>13.18006</td>
<td>18.2916</td>
<td>42.5025</td>
<td>22.0321</td>
<td>7.73621</td>
</tr>
</tbody>
</table>

F(2,16)=12.723, p<0.01. Follow-up paired t-tests showed significant effects between each pair: tone 2 and tone 3 with t(19)=-2.685 and p=0.015, tone 3 and tone 4 with t(19)=5.162 and p<0.001, tone 2 and tone 4 with t(19)=4.725 and p<0.001.

The analysis of variance also showed a significant effect of the interaction between tone and context, F(2,16)=6.509, p=0.009, η²=0.069. As a follow-up to the interaction, simple effects tests were performed on difference in tones at each contextual condition, and difference in context of each tone.

Paired t-tests on each tone between the two contexts showed a significant effect on tone 2 exclusively, with high offset context (M=2.532, SD=0.302), low offset context (M=2.849, SD=0.354), t(9)=-3.038, p=0.014. The mean of low offset context being significantly higher than the mean of high offset context is to say, for tone 2, different contexts result in significantly different task difficulty. Listeners find tone 2 preceded by a tone 4 harder to identify than preceded by a tone 1.

Two one-way repeated measures analysis of variance were run on within-variable tone
in each context condition, and both revealed significant effects. For high offset context, F(2, 27)=12.14, p<0.01, Tukey post hoc tests indicated significant effects between tone 2 (M=2.532, SD=0.302) and tone 3 (M=3.267, SD=0.844), p=0.034, between tone 3 and tone 4 (M=1.910, SD=0.585), p<0.01. For low offset context, F(2, 27)=13.2, p<0.01, Tukey post hoc tests indicated significant effects between tone 2 and tone 4, p<0.01 and tone 3 and tone 4, p<0.01.
These results indicate that in high offset preceding context, listeners find tone 3 noticeably harder to recognize than tone 2 and tone 4, while in low offset preceding context, they find both tone 2 and tone 3 noticeably harder than tone 4. This to some extent corresponds to the fact that the difficulty of identifying tone 2 varies in different contexts.
4 Discussion

4.1 Coarticulation and tone perception

Previous studies have observed that L2 listeners react differently to coarticulation effects than native listeners. Lee et al. (2008, 2010a) observed that while native listeners showed no difference in different contexts, L2 listeners’ performance decreased especially when the preceding tone and the target tone were not continuous in F0, or in other words, in conflicting context. This suggests that L2 listeners were not as good as native listeners at compensating for the variation of tone contour caused by coarticulation. From our results, we can see a difference between native and L2 listeners in the use of preceding context as well, especially in tone 2 with low offset context and tone 3 with high offset context. A tone 3 with high offset context was regarded as a conflicting context, while the other was not (Shih, 1988; Xu, 1994a,b, 1997).

In our experiments, all coarticulations observable were carry-over effects, which are the influence of the preceding context syllable “zhú” on the target syllable “wéi”. The context syllable “zhú” has either a high offset (tone 1) or a low offset (tone 4). According to Xu (1994b, 1997), carry-over effects are assimilatory. Specifically in our experiment, the onset of “wéi” becomes higher when it follows the high offset, and lower when it follows the low offset, as can be seen from Figure 2, which shows that the part of the contours corresponding to the last syllable in light gray are higher than those in dark gray in general, due to the higher offset of the preceding tone 1.

We are especially interested in two cases: tone 2 with low offset context (zh` u w´ ei) and tone 3 with high offset context (zh¯ u wˇ ei). The two cases showed the most salient difference between the two participant groups in the use of preceding context. The questions to be answered are: why did native listeners use the acoustic information of the preceding context while L2 listeners did not, and why did no such effects show in any other stimuli.
In Shen and Lin (1991), participants tended to predict tone 3 if the falling initial is steeper, the F0 difference is more noticeable and the distance from the onset to the turning point takes a larger proportion of the syllable. Jongman and Moore (2000) showed that listeners tend to select tone 3 when the precursor has higher pitch or is faster, and tone 2 if the opposite. Since carry-over effects are assimilatory, tone 2 following a lower offset starts with a lower onset, and the slope from the onset to the turning point is comparatively flat. All of these result in a similarity to tone 3. A higher offset preceding tone 3 might cause the opposite effect, raising the onset, increasing the steepness, and making tone 3 more similar to tone 2. Additionally, we used the half tone 3 in our tone 3 stimuli. The shorter duration might also shift the responses toward to tone 2 to some extent. Blicher et al. (1990) observed that listeners tend to perceive tone 2 as shorter and tone 3 as longer.

On the other hand, even though tone 4 is also influenced by coarticulation effects, there is no other tone in Mandarin that has a similar contour shape. The unique falling contour and the similarity of that contour to English declarative intonation makes its identification less challenging. In other words, the preceding contexts of tone 4 do not play an important role in its identification, because the difficulty of identifying tone 4 in different contextual conditions does not vary much. To confirm the assumption that the differences in importance area were due to different task difficulty, we further analyzed the final BPS data in Section 4.2.

4.2 Difference in task difficulty

The statistical analysis of final BPS reveals that listeners find the stimuli not equally difficult in identification tasks. There are significant effects on the interaction of tone and context.

First, only tone 2 shows significant effect by context. This is to say, listeners find tone 2 preceded by a low offset significantly harder than a high offset. This difficulty is mainly
due to the coarticulation effect and as a result tone 2 is more easily confused with tone 3. In most previous studies, tone 2 was found to be the hardest to identify (Gottfried and Suiter 1997, Lee et al. 2008, 2010a).

Second, in low offset context, both tones 2 and 3 are found to be significantly harder to recognize than tone 4, while in high offset context, tone 3 is found to be significantly harder to recognize than tones 2 and 4. This is also consistent to the fact that context showed significant effects only for tone 2. Tone 4 is the easiest among all. This was demonstrated by previous studies. Lee et al. (2010a) found that the identification of tone 4 was very robust, even based on only syllable onset information. Broselow et al. (1987) and Wang et al. (2006) pointed out that tone 4 presented in the final position is easy to identify for American listeners, because the this falling pitch contour in the final position is acoustically similar to the declarative in English.

However, even though tone 3 could be similarly influenced by coarticulation, there is no significant context effect. One possible explanation for this discrepancy is that, since we used the half tone 3, there is no real turning point, and the syllable duration is noticeably shorter than all other tones. The shorter duration does not give much acoustic information, which results in the finding that instead of tone 2, tone 3 is the hardest to identify in this experiment. Liu and Samuel (2004) found that when F0 information is insufficient, listeners use secondary cue such as duration. In the native importance maps of tone 3, we observe a high frequency area extending from approximately 1 kHz to 4 kHz and covering the target tone as well as some silence before and after it. The regions suggest the use of duration cue.

The statistical results did not show any significant effect of nativeness on final BPS. However, we still observed some differences in the point-biserial correlation coefficients. The lack of significant effect of nativeness could possibly be due to the small number of participants and the limited amount of data. It could also be due to the exclusion of subjects who could not sufficiently identify the speech in noise.
4.3 Summary

Shih (1988) based on the onset and offset pitch, categorized the four tones as “high-high”, “low-high”, “low-low” and “high-low”. This is a binary classification, and “mid” was not used. Xu (1994a,b, 1997) adopted this binary categorization, and defined “tone 1-2” and “tone 1-3” as a conflicting context, where a high offset is followed by a low onset, and “tone 4-2” and “tone 4-3” as a compatible context, where both the preceding offset and the following onset are low. They concluded that the amount of deviation of a tone from its canonical form is greater in a conflicting context. However, in our experiment, we observed a more severe deviation of tone 2 with low offset context. The final BPS also showed that tone 2 with low offset preceding context was regarded as more difficult by listeners. A low preceding context distorts the tone 2 contour comparatively severely and makes it more confusable with tone 3. As such, it is dubious to categorize tone 2 as “low-high”.

Our results indicate that the onset of tone 2 constructs a conflicting context with a preceding low offset and a compatible context with a preceding high context. This is more similar to a “high” pitch than a “low” pitch. However, we cannot categorize tone 2 as “high-high” either, considering that tone 2 has a clear rising contour. As such, the optimum solution would be “mid-high”.

Even though the final BPS showed no significant effects of nativeness, by comparing the importance maps we see a difference in the use of context between the groups. This happened with tone 2 with low offset context and tone 3 with high offset context, the two most challenging sentences according to final BPS statistics. That is to say, compared to L2 listeners, native speakers are better at using preceding context to facilitate their tone perception in adverse conditions. Native speakers better compensate for the tone contour distortion caused by coarticulation effects. The different abilities of recognizing tone variation might be one of the main reasons that native listeners outperform L2 listeners in harder
listening tasks. In easier tasks such as tone 4 identification, the two groups did not show much discrepancy.
5 Conclusion

This thesis investigated the following questions:

1. Which acoustic cues are used by native listeners when they identify Mandarin tones in a noisy environment?

2. Which acoustic cues are used by L2 listeners when they identify Mandarin tones in a noisy environment?

3. Do listeners with different language backgrounds use different acoustic cues for tone perception?

4. Do listeners use different cues for different tones and in different contexts?

The importance maps (Figure 5, 6) revealed that listeners of both groups use similar cues to identify tone 4. However, for tones 2 and 3, native listeners showed more use of preceding context, especially for tone 2 in low offset context and tone 3 in high offset context. By comparing the point-biserial correlations, we found significant differences between the two groups mainly in these two sentences as well (Figure 7). In both cases, tone contours are severely affected by coarticulation, and native listeners are able to better compensate for that change. These results might have pedagogical significance as well. Training L2 listeners to better make use of the acoustic cues in these significant area might improve their proficiency in tone perception.

The statistical analysis of final BPS did not show significant effects of nativeness, although, it did show significant effects of the interaction of tone and context. Specifically, tone 2 was significantly influenced by context. Tone 2 in low offset context was significantly harder for listeners than in high offset context, because lower preceding offset increases the possibility of confusing tone 2 with tone 3. In both contexts, tone 4 was the easiest to identify, while tone 3 was the hardest. This is different from previous research that found tone
2 to be the hardest (Gottfried and Suiter, 1997; Lee et al., 2008, 2010a), and we attribute this to our adoption of the half tone 3 and the fact that half tone 3 has shorter duration.

5.1 Future work

These results are based on the experiment run on native Mandarin listeners and L2 listeners who are novice Mandarin learners and have no other tonal language background. From these results we cannot be sure if the discrepancy in the use of context we observed is mainly caused by their language background or their Mandarin level. Comparing L2 listeners of different Mandarin level or comparing L2 listeners from both tonal and non-tonal language background could shed light on this question. Apart from tone coarticulation effects, how other segmental features affect tone perception in noisy environment should be worth studying as well. Xu and Xu (2003) explored how consonant aspiration affects the tone perception of clear speech. Different features of the consonant might affect the acoustic cues listeners use in tone perception.

In order to study how carry-over effects influence tone perception, in this experiment, we controlled the preceding context and asked the participants to identify the following tone. In future experiments, we will use following context, which might provide more cues to the identity of the preceding tone through coarticulation.
A Spectrograms, importance maps, and point-biserial correlations

Table 4: Native listeners results

<table>
<thead>
<tr>
<th>Spectrogram</th>
<th>Importance map</th>
<th>Correlation</th>
<th>Significant correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>tā jiào zhū wèi</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tā jiào zhù wèi</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tā jiào zhū wèi</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tā jiào zhù wèi</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tā jiào zhū wèi</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Continued below)
Table 4 (continued)

<table>
<thead>
<tr>
<th>Spectrogram</th>
<th>Importance map</th>
<th>Correlation</th>
<th>Significant correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Spectrogram" /></td>
<td><img src="image2" alt="Importance map" /></td>
<td><img src="image3" alt="Correlation" /></td>
<td><img src="image4" alt="Significant correlation" /></td>
</tr>
</tbody>
</table>

tā jiào zhù wèi
Table 5: L2 listeners results

<table>
<thead>
<tr>
<th>Spectrogram</th>
<th>Importance map</th>
<th>Correlation</th>
<th>Significant correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Spectrogram" /></td>
<td><img src="image2" alt="Importance map" /></td>
<td><img src="image3" alt="Correlation" /></td>
<td><img src="image4" alt="Significant correlation" /></td>
</tr>
<tr>
<td><img src="image5" alt="Spectrogram" /></td>
<td><img src="image6" alt="Importance map" /></td>
<td><img src="image7" alt="Correlation" /></td>
<td><img src="image8" alt="Significant correlation" /></td>
</tr>
<tr>
<td><img src="image9" alt="Spectrogram" /></td>
<td><img src="image10" alt="Importance map" /></td>
<td><img src="image11" alt="Correlation" /></td>
<td><img src="image12" alt="Significant correlation" /></td>
</tr>
<tr>
<td><img src="image13" alt="Spectrogram" /></td>
<td><img src="image14" alt="Importance map" /></td>
<td><img src="image15" alt="Correlation" /></td>
<td><img src="image16" alt="Significant correlation" /></td>
</tr>
<tr>
<td><img src="image17" alt="Spectrogram" /></td>
<td><img src="image18" alt="Importance map" /></td>
<td><img src="image19" alt="Correlation" /></td>
<td><img src="image20" alt="Significant correlation" /></td>
</tr>
</tbody>
</table>

(Continued below)
<table>
<thead>
<tr>
<th>Spectrogram</th>
<th>Importance map</th>
<th>Correlation</th>
<th>Significant correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Spectrogram" /></td>
<td><img src="image2.png" alt="Importance map" /></td>
<td><img src="image3.png" alt="Correlation" /></td>
<td><img src="image4.png" alt="Significant correlation" /></td>
</tr>
</tbody>
</table>

*Table 5 (continued)*

`tā jiào zhù wèi`
References


Ministry of Education of the People’s Republic of China (1958). 汉语拼音方案 [Scheme for the Chinese Phonetic Alphabet].


