Mandarin Tone 2/3 confusion revisited

Zixuan Li[1], *Zeyu Xie*[2], *Shijing Yu*[2], *and Yi Xu*[2]

[1]Department of Linguistics, University of Ottawa, Ottawa, Canada
[2]Speech, Hearing and Phonetic Sciences, University College London, London, United Kingdom
Zli073@uottawa.ca, yi.xu@ucl.ac.uk

Abstract

It is almost common knowledge that Mandarin Tone 3, especially in its citation form, is easily confused with Tone 2. But a review of the literature shows that the origin of this knowledge lies entirely in a single early study. In the present study, we recorded monosyllabic Mandarin words in all four tones either in isolation or in short lists. Acoustic analysis showed that isolated Tone 3 syllables mostly had fall-rise pitch contours, while those in lists often had low-fall contours. Also, the likelihood of Tone 3 showing fall-rise contours is associated with longer durations. A perception experiment with native Mandarin listeners further found that there was no more confusion between Tone 2 and Tone 3 than between any other two tones when the original syllable duration remained intact. When syllable duration was time-normalized, Tone 2 and Tone 3 indeed became more confusable. The finding may also raise questions about the nature of categorical perception, as naturally produced speech utterances may already consist of clearly separated phonetic categories, rather than being transformed into categories by speech perception from purely gradient phonetic variants. This is further supported by a corpus analysis, which showed that in pre-pausal positions virtually all occurrences of Tone 3 are of the low-fall variant with a short duration.

Index Terms: Tone 2/3 confusion, tone perception, time-normalization, categorical perception

1. Introduction

The earliest study reporting a high perceptual confusability of Mandarin Tone 2 and Tone 3 is Chuang and Hiki (1972) [1]. The study recorded monosyllabic, disyllabic and trisyllabic words produced by two male speakers and one female speaker from Taiwan, and asked four Chinese listeners (2 males and 2 females, place of birth and childhood unspecified) to name the tone of each syllable. For the monosyllabic words, the greatest number of confusions were between Tone 3 and Tone 2. Since then, no further studies have been conducted to replicate the high Tone 2/3 confusion in isolated monosyllabic Mandarin words by native listeners, although there are reports of high confusions by L2 learners [2]. Chuang and Hiki also suggested that the high confusion was due to the resemblance of pitch patterns of Tone 2 and Tone 3 [1]. Both the reported high confusion and the suggested likely reason have since been the premise for many subsequent studies of tone perception, most of which followed the categorical perception paradigm. In this paradigm, one acoustic dimension is made into an evenly spaced continuum, while all other acoustic aspects are kept constant [3]. For Tone 2 and Tone 3, the common practice is to keep the duration constant while changing f_0 contour into various continua [4][5][6][7][8]. This practice, however, has overlooked not only the large differences in duration across the four Mandarin tones already shown in [1], but also the fact that Tone 3 in Mandarin has multiple allophonic variants [9]. One of the reasons for this neglect is the unclarity about the distribution of the Tone 3 variants and their occurrence conditions. The present study is an examination of the relation between syllable duration and the likelihood of the fall-rise and low-fall variants of Tone 3 in speech production, and the effect of syllable duration on the perceptual confusability of Tone 2 and Tone 3.

1.1. The three variants of tone 3

Of the four lexical tones in standard Mandarin Chinese, Tone 1 (high-level), Tone 2 (rise), Tone 3 (fall-rise), and Tone 4 (fall) [9], Tone 3 exhibits the greatest degree of variability. It is known to have at least three distinct forms: fall-rise [214], low-fall [21], and rise [35] [9][10][11]. The rise variant is also known as the Sandhi Tone 3 which occurs only before another Tone 3.

The fall-rise variant of Tone 3, also known as the full or canonical Tone 3, is its citation form. That is, it is the form produced by native speakers when asked to say a monosyllabic Tone 3 word in isolation [1][11][12]. The prevalence of full Tone 3 was very high in citation forms as we found out in a pilot test for the current study, which failed to record any low-fall variant of Tone 3 from most speakers. This propensity is thus part of the Tone 2/3 confusion puzzle that this study hopes to address.

The low-fall variant of Tone 3 is the form that occurs when the tone is followed by any tone other than another Tone 3 [9][11][12]. It may also occur in a pre-pausal position [13], but the exact condition of its occurrence is unclear. Thus an additional objective of this study is to identify the conditions under which the low-fall variant of Tone 3 is likely to occur.

1.2. Distribution of Tone 3 variants as a function of syllable duration

As observed in the classical study of Chuang and Hiki [1] as well as subsequent studies [11][14][15][16][17], the fall-rise variant of Tone 3 is consistently longer in duration than all other tones. It is not clear, however, whether it is also longer than the low-fall variant of Tone 3 in the same context. Assuming that it is, the common practice of normalizing duration in the studies of categorical perception of the Tone 2/3 contrast could be problematic. This is because findings based on time-normalized stimuli cannot be properly interpreted, given that the short tone stimuli with rapid fall-rise contours are rarely encountered by listeners in natural speech.

One study did take duration into consideration by directly examining the effect of duration on the perceptual boundaries

of Tone 2 and Tone 3 [18]. They synthesized two groups of syllables with gradiently varied f_0 trajectories from a prototypical Tone 2 contour to a prototypical fall-rise variant of Tone 3 with both short (350 ms) and long (450 ms) durations. The results showed that the longer duration series elicited slightly more Tone 3 responses, thus demonstrating duration as a cue for distinguishing the two tones. However, by keeping the fall-rise variant of Tone 3 constant across both duration conditions, the study again leaves the perception results uninterpretable, as the combination of short syllable with fall-rise f_0 contour is again unlikely to occur in natural speech.

1.3. The present study

Given the inadequate knowledge about the true distinctiveness of Tone 2 and Tone 3 in Mandarin as reviewed above, the present study aims to achieve four research goals:

1. To elicit both fall-rise and low-fall variants of Tone 3 in well-separated monosyllabic words.
2. To examine if the distribution of the two variants is related to duration.
3. To examine the perceptual confusability of Tone 2/3 and the effect of time-normalization on their confusion.
4. To reexamine, through a corpus analysis, the assumption that the fall-rise variant is the default form of Tone 3 in pre-pausal positions.

2. Methodology

2.1. Experiment 1—Production

This experiment is designed for research goals 1 and 2. The first goal was tricky to achieve, because speakers had a strong natural tendency to say monosyllabic Tone 3 words in the citation form. But we eventually discovered a method that could easily elicit both the fall-rise and low-fall variants of Tone 3 as described next.

2.1.1 Stimuli

Two sets of minimal pairs covering the four Mandarin tones were constructed using the syllables /pa/ and /ma/. The selection of syllables followed two criteria:

(1) Each tone has a corresponding lexical item.

(2) All lexical items are commonly used high-frequency words.

Table 1: *Stimuli of Experiment 1*

	Tone 1	**Tone 2**	**Tone 3**	**Tone 4**
/pa/	八 (eight)	拔 (pull)	把 (hold)	爸 (dad)
/ma/	妈 (mom)	麻 (hemp)	马 (horse)	骂 (scold)

Two conditions were set in Experiment 1. In condition A, the stimuli were randomly grouped into a list of five characters, separated by commas. This setting was found to allow elicitation of low-fall variants of Tone 3 separated by silent pauses. A JavaScript displayed the stimuli on the computer screen, repeating each syllable 5 times in blocked random order. The order of the lists differed across the speakers. In condition B, each syllable was presented separately on the screen, again by a JavaScript, in a blocked random order that differed across speakers.

2.1.2 Speakers and recording procedure

25 students and teachers (14 females and 11 males) who were native Mandarin speakers born and raised in Beijing participated in the recording task. They were recruited from high schools and universities in Beijing. None of them reported having any speech or hearing disorders.

Recording was conducted online over Zoom, with the "original sound for musicians" option turned on. Participants were asked to do the recordings in a small quiet room with minimal reverberation. They were instructed to read aloud the Chinese characters displayed on the screen at a normal speech rate. For all speakers, the word lists (condition A) were recorded first. There was a short break between condition A and condition B.

2.1.4 Acoustic analysis

The recordings were first segmented into a total of 2000 sound files: 8 (syllables) x 5 (repetitions) x 2 (conditions) x 25 (speakers). These files were then annotated with ProsodyPro [21]. The syllable boundaries were manually marked based on the waveform, spectrogram and auditory judgment. The f_0 contours were generated based on Praat's automatic vocal pulse marking which was manual rectified for apparent errors. ProsodyPro then automatically extracted the following measurements from all the labeled intervals.

1. Duration.
2. minF0_loc_ratio —The relative position of the lowest f0 turning point, in terms of proportion to the total duration of the interval.

2.2. Experiment 2—Perception

Experiment 2 was designed to assess the confusability of all four Mandarin tones (research goal 3) by answering the following questions:

1. Is there a high Tone 2/3 confusion as previously reported?
2. Does time normalization affect perceptual tone confusion?
3. Which of the two Tone 3 variants is more confusable with Tone 2 and the other tones: fall-rise or low-fall?

2.2.1 Stimuli

From the recordings obtained in Experiment 1, 302 utterances (75 Tone 1, 75 Tone 2, 77 Tone 3, and 75 Tone 4 samples, with a nearly equal number of items in both syllable /pa/ and /ma/) were randomly selected for the perception test. In addition to the original sound files, another set was created by normalizing all syllable durations to 350 ms with a Praat script. The amplitude of all the samples was also normalized by peak-scaling to 0.99 in Praat.

Tone 3 was further divided into two groups: one with low tone 3 (21) and the other with full tone 3 (214), with nearly equal number of items in both groups. The distinction between the low tone 3 and full tone 3 was determined by the first two authors based on f_0 contours and auditory judgments.

2.2.2 Listeners and procedure

55 native Mandarin speakers participated in the perception experiment. All were born in northern China, speaking standard Mandarin. They were recruited from high schools and universities in Beijing. None of them reported any speech or hearing disorders.

The perception test was conducted online using the 'Homework' function of v8.chaoxing.com (a Chinese education website used by most Chinese university students). The participants were presented with 302 topics and were instructed to listen to the audio and choose the Chinese characters from the four options provided on screen. The characters were 妈 (Tone 1), 麻 (Tone 2), 马 (Tone 3) and 骂 (Tone 4) for syllable /ma/, and 八 (Tone 1), 拔 (Tone 2), 把 (Tone 3) and 爸 (Tone 4) for syllable /pa/. Each audio could be replayed multiple times by the listener, and there was no time limit for finishing a trial. The topics were presented in random order, and a different order was used for each listener. In total, 1,6603 responses = 302 (stimuli) x 55 (participants) – 7 (topics missed by the participants) were recorded.

2.3. Experiment 3—Corpus analysis

The corpus analysis was designed to reassess the assumption that the fall-rise variant of Tone 3 is its default form in pre-pausal positions [10][14]. The aim was to further broaden our knowledge about the distribution of the fall-rise and low-fall variants of Tone 3 given the observation of low-fall variant in sentence-final positions [13]. The corpus used is ASCCD (Speech Corpus of Chinese Discourse), which consists of 18 short essays read aloud by ten native speakers of Mandarin (5 females and 5 males) [20]. The corpus was phonetically and prosodically annotated. In the experiment we first identified all the Tone 3 syllables in pre-pausal positions, i.e., those before a period or a comma. We then measured their durations and analyzed their f_0 contours to determine whether they belong to the fall-rise or low-fall variant of Tone 3.

3. Results

3.1. Distribution of fall-rise and low-fall variants of Tone 3 and their relation to syllable duration

From Experiment 1, we were able to elicit many tokens of the low-fall variants of Tone 3 in the list condition as illustrated in Figure 1 left, together with many fall-rise variants (Figure 1 right).

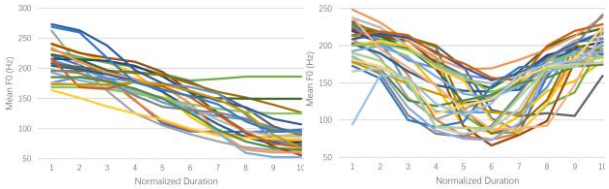

Figure 1: *Illustration of the low-fall and fall-rise variants of Tone 3. Left: 26 contours with short durations (<250 ms). Right: 30 contours with long durations (>400ms).*

To see if there is a relation between syllable duration and the likelihood of a fall-rise pattern in tone 3, Figure 2 shows a scatter plot of minF0_loc_ratio—low f_0 turning point relative to the syllable interval, as a function of syllable duration. A negative relation between syllable duration and the low turning point can be seen. That is, the longer the syllable, the earlier the low f_0 turning point relative to the full interval of the syllable.

There is a concentration of points with a minF0_loc_ratio of 1, which are from tokens where f_0 minima were detected exactly at the end of the syllable. This cluster may have biased the relation represented by the regression line. To reduce the bias, data points with minF0_loc_ratio = 1 were excluded in Figure 3.

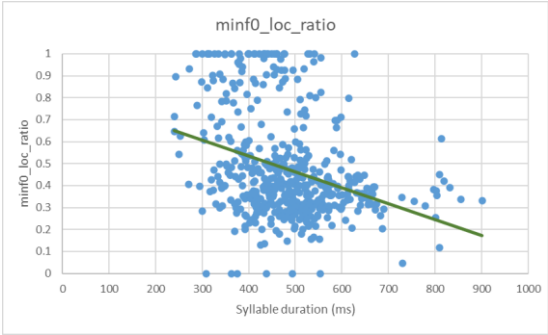

Figure 2: *Scatter plot of location of low f_0 turning point relative to syllable onset as a function of syllable duration.*

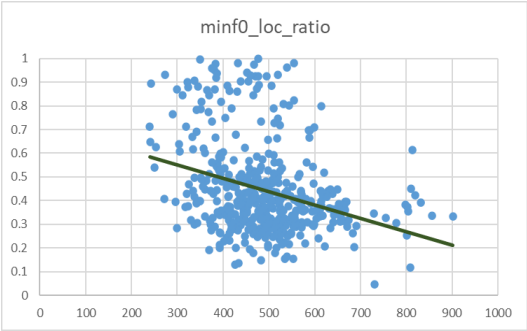

Figure 3: *Scatter plot of low f_0 turning point location relative to syllable interval as a function of syllable duration, without points where minF0_loc_ratio = 1.*

A Spearman's rank correlation analysis was done to examine the correlations between all pairs of measurements in Experiment 1, and the results are shown in Table 2. There is a significant negative correlation between syllable duration and f_0 turning point location in both conditions (all data, no ceiling effect controlled). That is, the longer the syllable duration, the earlier the f_0 minimum is likely to occur in the syllable. This indicates that the fall-rise variant of Tone 3 is more likely to occur when the syllable duration is longer.

Table 2: Results of Spearman's rank correlation

Minf0_loc / duration	$\rho = -0.32$	S = 27412573	$p < 0.001$
minf0_loc / duration (No ceiling effect)	$\rho = -0.29$	S = 21193568	$p < 0.001$

3.2. Perceptual distinctness of Tone 2 vs. Tone 3 and effect of time-normalization

Table 3 displays the confusion matrix of the four Mandarin tones in the original duration in Experiment 2. As can be seen, although Tone 3 has a lower overall recognition rate (89%) than the other tones, it is the most confusable with Tone 1 (4.07%) rather than Tone 2 (3.45%), and the confusion rate with Tone 4 (3.01%) is nearly as high. Tone 2 is more confusable with Tone 3 (1.53%) than other tones. But with its total recognition rate of 97.80%, it is hard to describe it as easily confusable with Tone 3.

Table 3: *Confusion matrix (%) of Mandarin tones in the original duration.*

Actual \ Perceived	Tone 1	Tone 2	Tone 3	Tone 4
Tone 1	**98.671**	0.098	0.098	1.132
Tone 2	0.334	**97.799**	1.531	0.334
Tone 3	4.067	3.445	**89.474**	3.014
Tone 4	0.404	0.354	0.008	**98.434**

Table 4 shows the confusion matrix of tones that have been normalized in duration. With the time-normalization, Tone 3 had the greatest reduction in recognition rate: from 87.47% to 80.05 (Table 3 vs. Table 4), while only Tone 1 showed a slight reduction from 98.67% to 97.22%. A one-way repeated-measures ANOVA with one within-subject factor—showed a significant effect of condition (Original vs. Time-normalized) on tone recognition rate ($F(1, 54) = 97.22$, $p < 0.001$).

Table 4: *Confusion matrix (%) of Mandarin tones in time-normalized durations.*

Actual \ Perceived	Tone 1	Tone 2	Tone 3	Tone 4
Tone 1	**97.221**	0.479	0.048	2.252
Tone 2	0.098	**98.329**	1.278	0.295
Tone 3	5.268	9.650	**80.047**	5.035
Tone 4	0.274	0.183	0.548	**98.995**

Table 5 shows the effect of time-normalization on the recognition rate of Tone 2 with either the low-fall and fall-rise variants of Tone 3. The recognition rate of the low-fall variant was much lower (82.27%) than that of the fall-rise variant (97.48%), indicating that it is the low-fall variant that had contributed the most to the overall low recognition of Tone 3 relative to other tones as seen in Table 3. Also, after time-normalization, it is the fall-rise variant of Tone 3 that reduced its recognition rate sharply from 97.48% to 84.12%, and the confusion is now mainly with Tone 2 (15.79%). In contrast, the low-fall variant of Tone 3 increased its confusion mainly with Tone 1 (10.27%) and Tone 4 (9.73%).

Table 5: *Confusion matrix of Tone 3 in original or time-normalized duration*

Actual \ Perceived	Tone 1	Tone 2	Tone 3	Tone 4
Original low-fall	7.73	4.27	**82.27**	5.73
Original fall-rise	0	2.53	**97.48**	0
Normalized low-fall	10.27	3.82	**76.18**	9.73
Normalized fall-rise	0	15.79	**84.12**	0.096

3.3. Distribution of fall-rise and low-fall variants of Tone 3 in ASCCD corpus and their relation to syllable duration

Finally, from the corpus analysis in Experiment 3, we identified 747 Tone 3 syllables in pre-pausal positions, 403 by female speakers and 344 by male speakers. Of the 747 tokens, only four presented fall-rise contours, rendering a rate of occurrence of 4/747 = 0.54% for the fall-rise variant of Tone 3 in pre-pausal positions in the corpus. The duration correlates of the Tone 3 variants are also consistent with what is found in Experiment 1. The mean duration of all the low-fall tokens was 245.8 ms, while that of the four fall-rising tokens was 365.3 ms. A caveat, however, is that most of the pre-pausal Tone 3 syllables in the corpus were not monosyllabic words, but part of disyllabic or trisyllabic words or phrases. This makes them non-equivalent to the monosyllabic words examined in Experiments 1 and 2.

4. Discussion and conclusion

The three experiments we have conducted were able to achieve the four research goals stated in 1.3. First, we discovered that by asking speakers to read aloud monosyllabic words in a list, the low-fall variant of Tone 3 could be readily elicited. Second, the low-fall variant of Tone 3 was associated with shorter syllable duration, while the fall-rise variant with longer syllable duration. Third, most significantly, Tone 2/3 confusion in the perception of the original utterances was *no higher* than confusions in any other tone pairs (Table 3); and among the two variants, it was the low-fall one that had greater confusions with other tones (Tone 1 and Tone 4, Table 5). Fourth, equally as significant, time-normalization of syllable duration increased the perceptual confusion of Tone 3 with Tone 2 (Table 4); and this particular effect was only on the fall-rise variant of Tone 3, while the increased confusion of the low-fall variant was with Tone 1 and Tone 4 rather than with Tone 2. Finally, the corpus analysis in Experiment 3 found that the fall-rise variant of Tone 3 occurred very rarely in pre-pausal positions: 4 out of 747 tokens, or 0.54%.

These results provide clear evidence, for the first time, against the widely held assumption that Tone 2 and Tone 3 in Mandarin are easily confusable. That assumption was based on only one perceptual study with naturally produced monosyllabic words in their original durations [1] which has never been replicated. The high Tone 2/3 confusion found in that study could be because of limited speakers (only 3) as well as listeners (only 4 native listeners), or, more likely, that the listening subjects were required to *name* the tone names, which is more difficult than to select the words in Chinese characters as done in the present study. Regardless of the real reason, however, a systematic replication of the early finding has been long overdue, and the results of the present study clearly demonstrate that the Tone 2/3 confusion has been an exaggeration or even a myth. This myth has been sustained or even reinforced over the years by studies that followed the categorical perception paradigm by using time-normalization to overlook not only the effect of duration, but also the natural correlation of duration with phonetically distinct variants of Tone 3.

The questions raised by the results of the present study may extend to the very premise of the theory of categorical perception, namely, that speech perception is endowed with a magical power to convert gradient acoustic variants into distinct categories. What we have found in this study is that naturally produced speech utterances may already be highly categorical in their acoustic signals, or articulatorily discontinuous, as recognized by the authors who proposed the original concept of categorical perception [22].

Finally, the current results by no means suggest that there is no potential difficulty in distinguishing between Tone 2 and Tone 3. As found in multiple studies, both children and L2 learners may indeed have a harder time with these two tones [22][24][25]. What the findings of the present study suggest is that, by the time native speakers have become mature language users, they are unlikely to be still struggling with the confusions between Tone 2 and Tone 3 as assumed and sustained by research based on the categorical perception paradigm.

5. References

[1] Chuang, C. K., Hiki, S., Sone, T., & Nimura, T. (1972). The acoustical features and perceptual cues of the four tones of standard colloquial Chinese. In Proceedings of the Seventh International Congress on Acoustics (Vol. 3, pp. 297–300). Akademiai Kiado.

[2] Wang, Y., Spence, M. M., Jongman, A. and Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America* **106**(6): 3649-3658.

[3] Liberman, A. M., Cooper, F. S., Harris, K. S., Hoffmann, H. S. and Griffith, B. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* **54**: 358-368.

[4] Wang, W. S.-Y. (1976). Language change. Annals of the New York Academy of Sciences, 280(1), 61–72. https://doi.org/10.1111/j.1749-6632.1976.tb25472.x

[5] Shen, X. S., & Lin, M. (1991). A perceptual study of Mandarin Tone 2 and Tone 3. Language and Speech, 34(2), 145–156.

[6] Zou, T., Zhang, J. and Cao, W. (2012). A comparative study of perception of tone 2 and tone 3 in Mandarin by native speakers and Japanese learners. In *Proceedings of 2012 8th International Symposium on Chinese Spoken Language Processing*. IEEE**:** 431-435.

[7] Wang, Y.-J. and Li, M.-J. (2010). The Effects of Tone Pattern and Register in Perceptions of Tone 2 and Tone 3 in Mandarin. *Acta Psychologica Sinica* **42**(09): 899-908.

[8] Dong, R. (2022). The Perception of Tone 2 and Tone 3 in Mandarin. In *Proceedings of 2021 International Conference on Social Development and Media Communication (SDMC 2021)*. Atlantis Press**:** 564-573.

[9] Chao, Y. R. (1968). A Grammar of Spoken Chinese. University of California Press.

[10] Duanmu, S. (2007). The Phonology of Standard Chinese (2nd ed.). Oxford University Press.

[11] Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics* **25**: 61-83.

[12] Shih, C. (1987). The phonetics of the Chinese tonal system, AT & T Bell Labs technical memo.

[13] Liu, F. and Xu, Y. (2005). Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica* **62**: 70-87.

[14] Ho, A. T. (1976). The acoustic variation of Mandarin tones. Phonetica, 33(5), 353–367. https://doi.org/10.1159/000259792

[15] Yang, J., Zhang, Y., Li, A. and Xu, L. (2017). On the Duration of Mandarin Tones. In *Proceedings of Interspeech***:** 1407-1411.

[16] Moore, B. C. and Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *Journal of the Acoustical Society of America* **102**: 1864-1877.

[17] Wu, Y., Adda-Decker, M. and Lamel, L. (2020). Mandarin lexical tones: a corpus-based study of word length, syllable position and prosodic position on duration. In *Proceedings of Interspeech 2020*. ISCA**:** 1908-1912.

[18] Blicher, D. L., Diehl, R. L., & Cohen, L. B. (1990). Effects of syllable duration on the perception of the Mandarin Tone 2/Tone 3 distinction: Evidence of auditory enhancement. Journal of Phonetics, 18, 37–49.

[19] Wu, Y., Adda-Decker, M., & Lamel, L. (2020). Mandarin lexical tones: A corpus-based study of word length, syllable position and prosodic position on duration. In Proceedings of Interspeech 2020 (pp. 1908–1912). ISCA. https://www.isca-speech.org/archive/Interspeech_2020/

[20] Li, A., Lin, M., Chen, X., Zu, Y., Sun, G., Hua, W., Yin, Z. and Yan, J. (2000). Speech corpus of Chinese discourse and the phonetic research. In *Proceedings of ICSLP2000.*

[21] Xu, Y. (2013). ProsodyPro — a tool for large-scale systematic prosody analysis (software/paper). Laboratoire Parole et Langage.

[22] Liberman, A., Harris, K. S., Eimas, P., Lisker, L. and Bastian, J. (1961). An effect of learning on speech perception: The discrimination of durations of silence with and without phonemic significance. *Language and Speech* **4**(4): 175-195.

[23] Li, C. N. and Thompson, S. A. (1977). The acquisition of tone in Mandarin-speaking children. *J. Child Lang.* **4**: 185-199.

[24] Hao, Y.-C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics* **40**(2): 269-279.

[25] Zou, T., Caspers, J., & Chen, Y. (2022). Perception of different tone contrasts at sub-lexical and lexical levels by Dutch learners of Mandarin Chinese. Frontiers in Psychology, 13, Article 891756. https://doi.org/10.3389/fpsyg.2022.891756