2024. *Proc Ling Soc Amer* 9(1). 5643. https://doi.org/10.3765/plsa.v9i1.5643.

Tonal identification in whispered speech

Ruyue Agnes Bi*

Abstract. This project aims to examine whether, and how, non-$F0$ cues facilitate the identification of lexical tones. A perception experiment is designed to explicitly test the impact of duration cues for Mandarin lexical tones when $F0$ is absent. We take a novel approach in which the secondary cue of interest is held constant, effectively controlling the type of information listeners receive. Future studies can potentially extend this methodology to examine other relevant cues, such as temporal envelope and intensity. The contribution of this paper is twofold: first, to propose an explanation for the inconsistent conclusions drawn in the literature on tonal identification in whispered speech; second, to devise a more well-controlled study shedding light on the nature of tonal perception.

Keywords. lexical tones; whispered speech; perceptual asymmetry; Biased Choice Model; Mandarin Chinese

1. Introduction. Phonetic contrasts are represented via multiple acoustic dimensions and signaled by multiple cues simultaneously. In the case of tones, it has long been established that the *fundamental frequency* ($F0$), defined as the number of cycles per second in a periodic waveform, serves as their primary cue (Gandour 1978; Yip 2002). Previous experimental studies have shown that even when all other cues are edited out of the auditory signals, native speakers of tonal languages can still reliably discriminate between various tones with near-ceiling accuracy (Abramson 1978). This suggests that $F0$ is a sufficient cue for tonal identification, raising the question of whether secondary cues serve any function in identification or simply come along for the ride. Researchers have mostly ruled out the null hypothesis that primary cues are necessary conditions for successfully identifying contrasts. Secondary cues, in fact, contribute substantially to the identification process, and listeners appear to be sensitive to a weighted combination of various acoustic dimensions (see, e.g., Di Paolo & Faber 1990; Wassink 2006; Zellou et al. 2020).

In this project, we are interested in determining the amount of information listeners can extract solely from secondary cues. Specifically, when listeners are deprived of a primary dimension of cues, how do they utilize the remaining acoustic features?

Since $F0$ cues overwhelmingly dominate in phonated speech, it is nearly impossible to detect the influences of other secondary cues, such as duration, amplitude contour, and vowel quality. However, this does not imply that listeners do not attend to these dimensions in tonal identification. Whispered speech provides an ideal context for our investigation. In whispers, $F0$ is absent since periodic voicing is replaced by a noise source. Unlike synthetic stimuli with superimposed average $F0$ contours, whispered speech occurs naturally in daily conversations. Using Mandarin as a case study, we propose a perception experiment to directly test the effect of one of the main secondary cues, namely **duration**, and to provide a preliminary answer to the following questions:

* I would like to thank Adam Albright, Athulya Aravind, Canaan Breiss, Martin Hackl, Jonah Katz, Michael Kenstowicz, audience of Phonology Circle at MIT, three reviews and audience at the LSA 2024 Annual Meeting for their constructive feedback. I would also like to thank Piper Oren at the MIT Behavioral Research Lab and Leo Rosenstein at the MIT Linguistics Experimental Syntax & Semantics Lab for their help in collecting the perception experiment data. Special thanks to Edward Flemming for extensive discussion on this project. All mistakes are mine.
Author: Ruyue Agnes Bi, Massachusetts Institute of Technology (ruyuebi@mit.edu).

- Can listeners successfully distinguish tones when $F0$ is absent?
- Does an above-the-average accuracy necessarily mean successful identification? Is there genuine sensitivity to tonal contrasts in whispers, or would perceptual bias alone largely account for the data?
- If there is genuine sensitivity, what are the secondary cues recruited by listeners in these inhibited conditions, and how are they utilized?

The rest of the paper is structured as follows: Section 2 provides the theoretical background on Mandarin lexical tones, reviews existing experimental studies on tone identification in whispered speech, and discusses some of their methodological shortcomings. Section 3 reanalyzes a set of data reported in an earlier paper using the Biased Choice Model, a statistical tool tailor-made to separate genuine sensitivity from biases. Section 4 describes the experimental designs and procedures. Section 5 reports the findings and discusses our analyses. Section 6 concludes.

2. Background.

2.1. MANDARIN TONES BASICS. Mandarin Chinese has four contrastive lexical tones: Tone 1 (high level 55), Tone 2 (rising 35), Tone 3 (low falling rising 214), and Tone 4 (falling 51). In citation forms of normal speech, Tone 3 has the longest duration and lowest average intensity, while Tone 4 has the shortest duration and highest average intensity (Chang & Yao 2007). The typical patterns of the four tones in citation form are illustrated in Figure 1, where each tone is plotted with its average duration proportional to that of Tone 3.

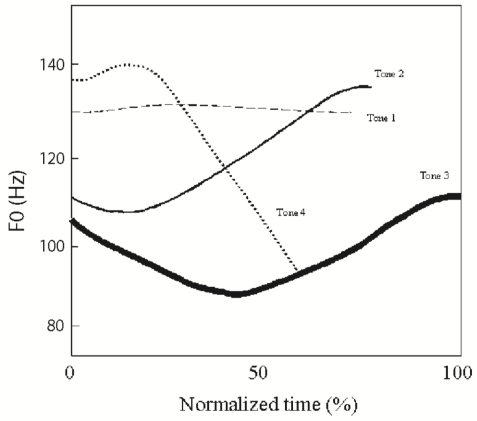

Figure 1. Typical $F0$ contours for the four contrastive tones in Mandarin (Liu & Samuel 2004)

However, it is important to note that this pattern does not necessarily hold in continuous speech. Recent production experiments and corpus studies have shown that the durational contrasts of the four lexical tones are much less consistent and less distinguishable in continuous speech. For instance, Yang et al. (2017) found that, compared to isolated monosyllables, the lexical tones in text-reading and conversational speech are significantly shorter and narrowly distributed within a smaller range. Specifically, in text-reading speech, Tone 2 is the longest and Tone 1 is the shortest, while in conversational speech, all four tones exhibit similar durations. Wu et al. (2020) carefully analyzed the duration patterns of lexical tones in fluent speech based on a corpus of approximately 220 hours of recordings. They concluded that although Tone 3 tends to have the longest duration in general, independent factors such as word length, syllable position, and prosodic position also play significant roles. Generally, the picture becomes much less clear

outside of citation forms. As we will see later, this inconsistency is reflected in the recorded stimuli for our perception experiment, raising questions about the reliability of duration differences as a perceptual cue in phonated speech.

2.2. EARLY WORK ON TONAL IDENTIFICATION IN WHISPERS. Previous experimental studies show a confusingly wide range of accuracy rates for tone identification in whispered speech: some have found performance hovering around chance (Abramson 1972; Miller 1961), while others have reported performance well above chance (Jensen 1958). However, a closer look into these studies reveals that we cannot simply take these results at face value.

Jensen (1958) is one of the earliest experimental studies of lexical tones in whispers. It had a rather minimal design: binary, pair-wise identification tasks were set up for **Mandarin**, while only one pair of tones each was tested for **Norwegian**, **Swedish**, and **Slovenian**. Note that this design can arbitrarily inflate accuracy rates since the other options are ruled out a priori for the participants. A very limited number of participants took part in the experiment, and only recognition scores (i.e., percent correct) data were reported in the paper.

Miller (1961) conducted several experiments to examine the identification of tonal contrasts in whispers for **Vietnamese**. Vietnamese has six lexical tones, which Miller refers to as *mid-level*, *low falling*, *high rising*, *high broken*, *low rising*, and *low dipped*. Three subjects participated in the experiments. A baseline experiment tested words in complete sentences, and participants performed exceptionally well, achieving nearly a 100% accuracy rate. The rest of the experiments tested pairwise identification, yielding mixed results. Experiment A tested mid level versus high rising, and 57% (compared to 50% by chance) of the tokens were correctly identified. Experiment B extended the tasks in A to the entire paradigm and tested 15 ($= (6 \times 5)/2$) pairs of tones. The accuracy rates varied from 37% to 58% (compared to 50% by chance). Experiment C tested all tonal variations of the same syllable, with all six options given as possible responses. The accuracy rates ranged from 22% to 47% (compared to 16.7% by chance). Miller provided an explicit breakdown of the identification data (summarized in Table 1), and pointed out that the high accuracy rate in identifying the low-dipped tone is not particularly convincing when considering that participants frequently misidentified other tones as low dipped. It is also worth noting the asymmetry in, for example, how often high rising is misidentified as low rising ($\approx 16.8\%$) versus how often low rising is misidentified as high rising ($\approx 2.5\%$).

Produced \ Response	*Mid Level*	*Low Falling*	*High Rising*	*High Broken*	*Low Rising*	*Low Dipped*	Total # Presented
Mid Level	41	28	21	13	16	19	138
Low Falling	15	83	3	3	15	21	140
High Rising	25	20	42	11	23	16	137
High Broken	3	3	15	63	34	35	153
Low Rising	18	25	3	34	29	10	119
Low Dipped	9	1	6	13	16	92	137

Table 1. Confusion matrix based on results from experiments B and C in Miller (1961)

Abramson (1972) discusses the identification of whispered **Thai** lexical tones both in isolation and in meaningful sentence contexts. Thai has five contrastive tones: *middle*, *low*, *falling*, *high*, and *rising*. The aggregated results of two sets of whispered monosyllabic words in isolation are reconstructed and shown in Table 2. His main takeaway was that secondary cues are redundant and only aid identification when tonal phonemes are embedded in "a sufficiently long

phonetic environment." He also claims that isolated monosyllabic words appear to be the most stringent test of the phonological distinctiveness of tonal features. However, this assertion is debatable. As evidenced by the case of Mandarin, duration information is, in fact, less distinctive in continuous speech, which presumably makes the identification of lexical tones in continuous speech at least as challenging as in citation forms.

Produced \ Response	*Mid*	*Low*	*High*	*Falling*	*Rising*	Total #
Mid	25	26	17	3	24	95
Low	22	51	3	9	10	95
High	28	32	23	5	7	95
Falling	30	20	11	34	0	95
Rising	12	31	7	5	40	95

Table 2. Confusion matrix reconstructed from whispered data as reported in Abramson (1972)

Liu & Samuel (2004) provides the first systematic study we found on **Mandarin** tone identification in environments with limited $F0$ information. The authors designed two experiments testing both synthetically processed stimuli and natural whispered ones. In the case of the former, the $F0$ contour of a phonated stimulus was extracted, and then it was resynthesized with a white noise source, while leaving the orginal phonetic, amplitude, and duration information intact. They found that Tone 3 stimuli were correctly identified at a level well above chance, Tone 1 identification was near chance, and Tone 2 and Tone 4 identification were somewhere in between. An acoustic analysis of the human whispered stimuli reflects a correlation between Tone 3 syllable duration and their perception.

	HUMAN WHISPERED				SYNTHETICALLY MANIPULATED			
Produced \ Response	*Tone 1*	*Tone 2*	*Tone 3*	*Tone 4*	*Tone 1*	*Tone 2*	*Tone 3*	*Tone 4*
Tone 1	18	25	20	37	12	36	25	27
Tone 2	11	34	37	18	8	44	32	16
Tone 3	6	17	68	9	7	27	58	8
Tone 4	25	15	19	41	11	30	30	29

Table 3. Response percentages for each tone reported in Liu & Samuel (2004)

As an interim summary, these pioneering studies have provided us with some important insights into the nature of tone identification; however, they also suffer from a few shortcomings that potentially explain their diverging results. First, they have mainly relied on percent correct to evaluate the success of identification. This metric can be deceptive and leaves many nuances of the data unexplored. A more advanced statistical model is needed to isolate independent bias from true perceptual distinctions. Secondly, earlier experiments tended to be smaller in scale, involving a limited set of stimuli and only a handful of participants. Larger-scale and more systematic tests are necessary to obtain a more complete picture. We aim to address both of these limitations in our study.

3. Modeling perceptual distances. In this section, we will introduce our statistical model of choice — the Luce's Biased Choice Model (Luce 1963; Estes 1997), by re-analyzing the results from Miller (1961) to arrive at a clearer conclusion.

Before moving on to a more complicated metric, let us first establish why percent correct is

inadequate for measuring confusability. Listeners may be predisposed to respond with a particular category, independent of the stimulus they actually hear, due to factors such as type or token frequencies. We categorize these factors under *response bias*. Response biases can artificially inflate the accuracy rate. For example, if a participant responds with Tone 3 to every stimulus they are given, they will identify Tone 3 correctly 100% of the time, but it does not necessarily mean that they are actually distinguishing Tone 3 from the others. Moreover, percent correct does not provide a complete description of subjects' behavior in perception tasks, especially regarding asymmetries in confusability. As seen in the confusion matrices from the last section, errors in these studies are generally not randomly distributed among the incorrect categories.

The Biased Choice Model (BCM) is designed to quantify perceptual distances through measures of confusability[1]. In theory, it factors out the effects of bias to recover *perceptual distance* and is applicable to multi-category identification data. The core assumption is that the probability of identifying a stimulus s_i as belonging to the response category r_j is proportional to the similarity between stimuli s_i and s_j, and the bias towards the response r_j. Formally,

$$P(r_j|s_i) \propto \eta_{ij} b_j, \text{ where } \eta_{ij} = \textbf{similarity} \text{ between } s_i \text{ and } s_j \\ b_j = \textbf{bias} \text{ towards response } r_j \tag{1}$$

Similarity coefficient η_{ij} ranges between 0 and 1: $\eta_{ij} = 1$ means that s_i and s_j are identical, while $\eta_{ij} = 0$ means they are infinitely distinct. Since every stimulus is identical to itself, $\eta_{ii} = 1$. Moreover, we assume similarity is symmetric; in other words, for each pair of stimuli s_i and s_j, $\eta_{ij} = \eta_{ji}$. Similarity coefficients are crucial in BCM, as they are used to calculate the perceptual distance between each pair of stimuli s_i and s_j:

$$\textbf{distance } d_{ij} \coloneqq -\log \eta_{ij} \tag{2}$$

Note that $d_{ii} = -\log 1 = 0$, i.e., there is zero distance between any stimulus and itself. When $\eta_{ij} = 0$, $d_{ij} = -\log 0 = -\infty$, i.e., a similarity score of 0 maps to infinite distance.

How often two stimuli s_i and s_j are confused with each other depends on their similarity. The difference in biases explains why s_i might be misidentified as s_j more or less often than s_j being misidentified as s_i. We want to find values for similarities η and biases b that yield the best fit between what we observe in the actual confusion matrix and what the model predicts for the count in each cell. That is to say, the model finds the coefficients that provide the best estimations for the number of times stimulus s_i is identified as response r_j, for every s_i. When implemented in R, BCM is reformulated as a log-linear model, but the core idea remains the same.

Returning to the data in Table (1) reconstructed from Miller (1961), we model the count in each cell of the confusion matrix, using *response*, *stimulus tone*, and *perceptual distance* as predictors. Additionally, as previously noted, we observe a suspiciously high accuracy rate for the Low Dipped tone. Therefore, a biased choice log-linear model is fitted in R, with the baseline for response set at Low Dipped. The model output allows us to calculate bias parameters for each tone and estimate the perceptual distance between each pair of tones.

All five response coefficients have $p < 0.001 \ll 0.05$. Hence, we can safely conclude that the biases towards the other five tones are significantly different from that towards Low Dipped.

[1] The discussion here is adapted from Flemming (2020) lecture notes, which cover the model in more depth and with case studies.

Specifically, the bias toward the Mid Level tone $b_{\text{ML}} \approx 0.067$, the bias toward the Low Falling tone $b_{\text{LF}} \approx 0.593$, the bias toward the High Rising tone $b_{\text{HR}} \approx 0.053$, the bias toward the High Broken tone $b_{\text{HB}} \approx 0.079$, the bias toward the Low Rising tone $b_{\text{LR}} \approx 0.071$, and the bias toward the Low Dipped tone $b_{\text{LD}} \approx 0.137$. Somewhat unexpectedly, there is an overwhelming bias towards Low Falling and only a moderate bias towards Low Dipped.

With biases factored out, we can now focus on identifying the perceptual discriminability between each pair of tones, which is reflected by the d values obtained in the model. The distances between each pair of tones, except for the one between High Broken and Low Rising, are all statistically significant ($p < 0.001 \ll 0.05$). This suggests that even in whispered speech, almost all pairs of tones are distinct enough from each other. Specifically, Low Dipped and Low Falling are the most distinct from each other ($d_{\text{LD, LF}} \approx 5.996$), immediately followed by High Broken and Low Falling ($d_{\text{HB, LF}} \approx 4.997$). High Rising and Mid Level are the least distinct ($d_{\text{HR, ML}} \approx 0.614$).

This analysis accounts for the mixed accuracy results reported in Miller (1961) and leads to a more definitive conclusion: Vietnamese tones are sufficiently distinct from one another even in whispered speech. This indicates that secondary cues alone can still provide a significant amount of information for tonal identification.

4. Perception experiment. A relatively recent study, Jiao & Xu (2019) (J&X), tested phonated versus whispered utterances of the full tonal paradigm in Mandarin. Participants were asked to identify individual words in citation forms. The stimuli consist of five sets of syllables composed of only vowels (/a/, /ɤ/, and /u/) or with glide onsets (/i/-[ji], /y/-[jy]).

Tone \ Vowel		/a/	/ɤ/	/i/	/u/	/y/
Tone 1	Character	啊	婀	衣	乌	迂
	Pinyin	ā	ē	yī	wū	yū
	Glossary	'oh'	'graceful'	'clothes'	'black'	'winding'
Tone 2	Character	啊	鹅	姨	无	鱼
	Pinyin	á	é	yí	wú	yú
	Glossary	'eh'	'goose'	'aunt'	'nothing'	'fish'
Tone 3	Character	啊	恶	椅	五	雨
	Pinyin	ǎ	ě	yǐ	wǔ	yǔ
	Glossary	'what'	'nausea'	'chair'	'five'	'rain'
Tone 4	Character	啊	饿	意	物	玉
	Pinyin	à	è	yì	wù	yù
	Glossary	'ah'	'hungry'	'meaning'	'thing'	'jade'

Table 4. List of syllables for the perception experiments in Jiao & Xu (2019)

Twenty-two native Mandarin speakers (average age: 20.2; 12 identifying as females) participated. The authors found that, unsurprisingly, tones are much more poorly identified in whispered than phonated utterances (49.9% vs. 96.4% correct). Nevertheless, based on the confusion matrices (see Table 5), they concluded that tone perception is better than chance in whispers but only speculated on what might have contributed to the process, leaving the questions of how the listeners achieve this level of success and what cues they are relying on.

As J&X point out, a likely candidate for such a secondary cue, at least in Mandarin[2], is dura-

[2] We suspect that the dominant secondary cue, when the primary one is absent, is idiosyncratic and differs from one tonal language to another.

	Target \ Response	*Tone 1*	*Tone 2*	*Tone 3*	*Tone 4*
PHONATED	*Tone 1*	98.48%	0.76%	0.38%	0.38%
	Tone 2	0.76%	98.86%	0.00%	0.38%
	Tone 3	0.00%	7.20%	92.42%	0.38%
	Tone 4	0.00%	0.38%	3.79%	95.83%
WHISPERED	*Tone 1*	23.86%	20.45%	8.33%	47.35%
	Tone 2	20.83%	31.06%	27.27%	20.83%
	Tone 3	0.76%	12.50%	84.47%	2.27%
	Tone 4	19.32%	12.12%	8.33%	60.23%

Table 5. Confusion matrices of tone identification rates for natural stimuli in Jiao & Xu (2019)

tion. Instead of drawing conclusions from post hoc statistical analyses, which are common practice in this literature, we will explicitly test the effect of duration by incorporating it into the experimental design.

4.1. DESIGN. We redesign the experiments conducted in Jiao & Xu (2019) to: (i) synthetically manipulate the durations of the target words, and (ii) embed the target words within carrier phrases. Embedding the target words in carrier phrases allows for more accurate measurements of duration, especially in whispers[3]. In most cases, we can identify clear spectrogram patterns of the target words that differ from those of their surrounding context.

The target words consist of five sets of monosyllables, each composed either solely of a vowel ([ɤ] and [u]), or with a glide or nasal onset (/i/-[ji], /y/-/[jy], and /a/-[ma]). These sets are identical to those used in J&X, except for the /a/ quadruplet. Different tonal versions of the syllable /a/ function as discourse markers with meanings such as 'ah' or 'what?' and they correspond to the same character 啊, making it challenging to present them as distinct options in a forced-choice task. The five sets of tone quadruplets are listed in Table 6.

Tone \ Vowel		/ma/	/ɤ/	/i/	/u/	/y/
Tone 1	Character	妈	婀	衣	乌	迂
	Pinyin	mā	ē	yī	wū	yū
	Glossary	'mother'	'graceful'	'clothes'	'black'	'winding'
Tone 2	Character	麻	鹅	姨	无	鱼
	Pinyin	má	é	yí	wú	yú
	Glossary	'hemp'	'goose'	'aunt'	'nothing'	'fish'
Tone 3	Character	马	恶	椅	五	雨
	Pinyin	mǎ	ě	yǐ	wǔ	yǔ
	Glossary	'horse'	'nausea'	'chair'	'five'	'rain'
Tone 4	Character	骂	饿	意	物	玉
	Pinyin	mà	è	yì	wù	yù
	Glossary	'scold'	'hungry'	'meaning'	'thing'	'jade'

Table 6. Target stimuli in the current study

These items are embedded in the two carrier phrases (1) and (2). In (1), the word preceding the target item, *shuō* 'say', ends at the upper half of the tonal space, while in (2), the second syllable

[3] In a previous iteration of this experiment, isolated syllables were used following J&X's original design. However, it turned out to be very difficult to establish consistent segmentation criteria for measuring syllable duration, as there is very little energy concentration registered on the spectrogram of the whispered tokens.

of the preceding word *chóngfù* 'repeat' ends at the lower half of the tonal space. Although we do not anticipate the immediately preceding environment to affect tone identification in this setup, it is controlled for in case any coarticulation effect does occur.

(1) 请说___ 给我听。
qǐng shuō ___ *gěi wǒ tīng.*
please say ___ give me listen
'Please say ___ to me.'

(2) 请重复___ 给我听。
qǐng chóngfù ___ *gěi wǒ tīng.*
please repeat ___ give me listen
'Please repeat ___ to me.'

Previous studies have mostly tested target words either in isolation or in coherent sentence contexts. We chose to use carrier phrases that do not provide semantic contextual information to avoid potential confounds.

The original stimuli were recorded by two female native speakers of Mandarin, who grew up in the eastern coastal region of China. From each speaker, 5 (syllables) $\times 4$ (tones) $\times 2$ (phonated vs. whispered registers) $\times 2$ (carrier phrases) $= 80$ stimuli were recorded. The recordings were made in a soundproof booth at a sampling rate of 44.1kHz as 16-bit format mono sound files.

4.1.1. SEGMENTATION CRITERIA AND ACOUSTIC ANALYSIS. Phonated target words in carrier phrases were segmented based primarily on waveform periodicity, excluding initial and final glottalization. If there is vowel coarticulation, we referred to intensity and pitch tracking to determine a reasonable boundary. Likewise, for whispered items, we searched for a stable spectrogram shape, i.e., a consistent energy concentration pattern, considering both the ending of the preceding sound and the beginning of the following one. All boundaries were set at the nearest zero-crossing for consistency.

An acoustic analysis was conducted to understand the duration distribution. As shown in Figure 2, whispered tokens tend to be longer than their phonated counterparts. In terms of duration differences between tones within each register, duration contrasts are largely neutralized among the phonated tokens, where all target words are of similar length. Duration contrasts among the whispered tokens show a more familiar pattern — Tone 3 tends to be the longest, and Tone 4 tends to be the shortest.

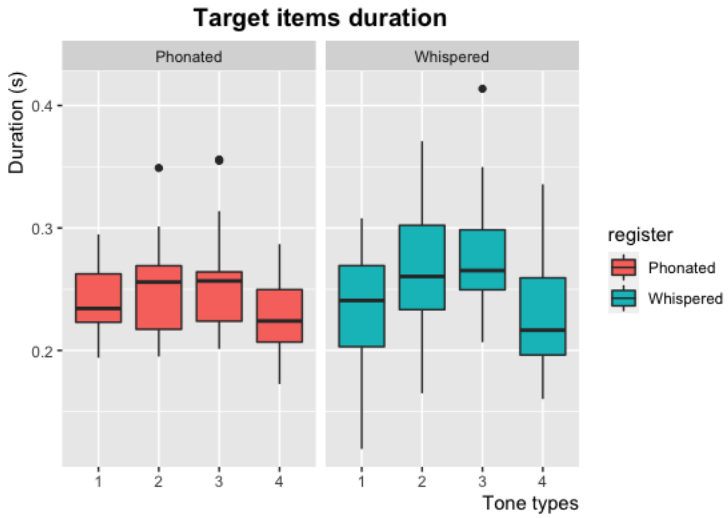

Figure 2. Duration distribution of the recorded items in phonated vs. whispered condition

Since the longest average duration in whispered speech is 305ms (whispered Tone 3), and the shortest is 192ms (whispered Tone 4), we created two sets of stimuli with target word durations normalized to these two values, respectively. Specifically, we ran a Praat script to manipulate the duration tier, either stretching or compressing the duration of each target word to a preset value while keeping the rest of the carrier phrase intact. The underlying assumption is that even with duration changes, tones remain distinct from each other, at least in the phonated register. We now have a LONG-DURATION set where the duration of each target word equals the longest average duration across the four tone categories in whispers (305ms in this case), and a SHORT-DURATION set where the duration of each target word equals the shortest average duration across the four tone categories in whispers (192ms in this case). This creates a total of 320 tokens, fully crossing 4 tones, 5 syllables, 2 carrier phrases, 2 registers, 2 speakers, and 2 duration levels.

The entire sentences, including the carrier portion, were played to participants in order to provide some clues regarding the overall speech rate of each speaker and where the boundaries for the target items lie.

4.1.2. PROCEDURE. Stimuli were presented over headphones. At each trial, after a sentence was played, participants were given four choices, each of which was a simplified Chinese character corresponding to one of the relevant tone quadruplets in orthography. Participants were instructed to choose the character that best matched what they heard. Each audio could be repeated at most once to prevent situations where participants might get distracted and accidentally miss the target word. Participants were encouraged to take a break halfway through the experiment. The entire experiment lasted no longer than 30 minutes.

4.2. PARTICIPANTS. Forty-one native Mandarin speakers between the ages of 20 and 41 (median: 28) participated in the experiment, with 28 identifying as females and 13 as males. None of the participants reported any hearing or speech impairments. They were compensated for their participation in the study.

4.3. HYPOTHESIS AND PREDICTIONS. Note that holding the duration of the target words constant allows us to control the amount of information listeners receive. This effectively **removes the duration cues** and should result in **lower accuracy rates across the board** if listeners rely on them to identify tones.

Furthermore, following Liu & Samuel (2004), we hypothesize that the longer a token is, the more likely it is for listeners to identify it as Tone 3. The two levels of the duration variable, LONG-DURATION and SHORT-DURATION, are intended to test this hypothesis. If it is borne out, we should observe **a greater bias towards Tone 3 in the LONG-DURATION condition compared to the SHORT-DURATION condition**.

5. Results and analyses.

5.1. DATA PRUNING. Excluding practice trials, a total of $13,284$ observations were collected. Data from the participant coded as "21" were discarded because they reported difficulty reading the simplified Chinese characters used to present response options. For the remaining $12,960$ data points, the reaction time data in seconds appears as in 3. 95% of the data points fall between 0.11s and 5.85s. Due to some extreme outliers, observations with reaction times greater than 8s, which comprises of approximately 1.86% of the total, are discarded.[4]

[4] It is worth noting that while setting accuracy rates for the phonated portion above a certain threshold could also be a reasonable exclusion criterion, we decided against it for two reasons: (i) since the phonated tokens are also syn-

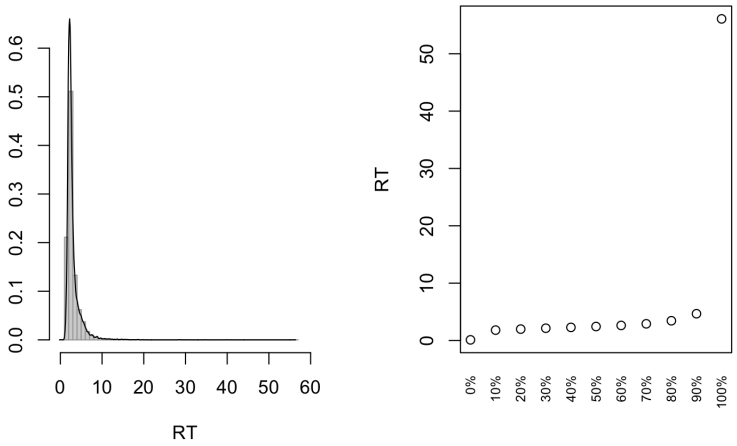

Figure 3. Reaction time distribution in the raw data

The reaction time data distribution after pruning, as shown in Figure 4, suggests that there is no significant difference between phonated and whispered items regarding how long it takes participants to make a decision.

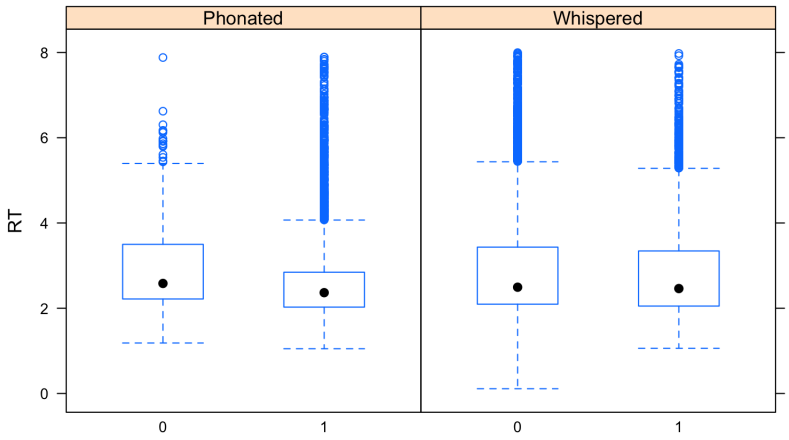

Figure 4. Reaction time distribution after pruning

5.2. ACCURACY & CONFUSION MATRIX. As expected, tone identification is much less accurate in whispered utterances compared to phonated ones (35.04% vs. 96.48% correct). Interestingly, in phonated utterances, the identification of Tone 2 is no longer at ceiling (90.04% for Tone 2, as opposed to 99.13% for Tone 1, 99.26% for Tone 3, and 97.41% for Tone 4.).[5] We did not

thetically manipulated, theoretically, we should not expect participants to perform as well as for natural stimuli; (ii) allowing noise in the data, which can often be informative in analyzing perception bias.

[5] In all cases where a phonated Tone 2 token was misidentified, around half were identified as Tone 1 and the other half as Tone 3. Mistaking Tone 2 for Tone 3 somewhat makes sense, as they have similar contours up until the turning point of the latter. However, mistaking Tone 2 for Tone 1 is more mysterious. Upon closer examination of these errors, more than half were made in the context of Speaker A, Carrier (2). These errors could be due to, rather than

expect the manipulation of the durations to affect tone identification for phonated speech, since F_0 should be a sufficient phonetic cue. Nonetheless, the fact that the accuracy rates for the other three tones are near-ceiling gives us some confidence the edited stimuli still sounded natural.[6]

Target \ Response	*T1*	*T2*	*T3*	*T4*
Tone 1	1601	8	4	2
Tone 2	73	1447	86	1
Tone 3	5	4	1617	3
Tone 4	11	9	22	1579

Target \ Response	*T1*	*T2*	*T3*	*T4*
Tone 1	99.13%	0.50%	0.25%	0.12%
Tone 2	4.54%	90.04%	5.35%	0.06%
Tone 3	0.31%	0.25%	99.26%	0.18%
Tone 4	0.68%	0.56%	1.36%	97.41%

Table 7. Aggregated confusion matrix for **phonated** synthetic stimuli

Among the whispered tokens, the accuracy rates for Tone 2 (46.42%) and Tone 3 (57.40%) are well above chance. However, as discussed in earlier sections, these results should not be taken at face value, as they might be mainly due to response biases defaulting to Tone 2 or Tone 3 when participants are uncertain. Indeed, when a Tone 1 stimulus is presented, participants are, on average, more likely to identify it as Tone 2 or Tone 3 than the actual type itself, and the same trend is observed for Tone 4 stimuli. The accuracy rates for Tone 1 (16.96%) and Tone 4 (19.37%) are below chance (25%). Compared to the data reported in J&X (23.86% accuracy rate for Tone 1, 31.06% for Tone 2, 84.47% for Tone 3, and 60.23% for Tone 4), we observe overall worse performance with edited whispered stimuli. Albeit indirectly, this lends some support to our first hypothesis.

Target \ Response	*T1*	*T2*	*T3*	*T4*
Tone 1	266	519	531	252
Tone 2	210	726	511	117
Tone 3	142	292	896	231
Tone 4	256	435	562	301

Target \ Response	*T1*	*T2*	*T3*	*T4*
Tone 1	16.96%	33.10%	33.86%	16.07%
Tone 2	13.43%	46.42%	32.67%	14.80%
Tone 3	9.10%	18.71%	57.40%	4.80%
Tone 4	16.47%	27.99%	36.16%	19.37%

Table 8. Aggregated confusion matrix for **whispered** synthetic stimuli

5.3. STATISTIC MODELING AND DISCUSSION. We fit a Biased Choice loglinear model in R to separate response bias from perceptual similarity and capture the asymmetries in errors. Recall that from the model output, we can calculate bias parameters for each tone and estimate the perceptual distance between each pair of them.

We found a significant main effect of **response**. Note that the baseline for the response is set at Tone 3, while the rest are sum coded. Since all three response coefficients have $p < 0.001 \ll 0.05$, we can safely conclude that the biases towards Tone 1 ($b_{\text{T1}} \approx 0.12$), Tone 2 ($b_{\text{T2}} \approx 0.33$), and Tone 4 ($b_{\text{T4}} \approx 0.12$) are significantly different from that towards Tone 3 ($b_{\text{T3}} \approx 0.42$). There is a large bias towards choosing Tone 3 and a smaller yet still significant one towards Tone 2.

We also found a significant main effect of **distance**. The distance measures between each tone pair, except for D_{T1T4}, are statistically significant (with $p \ll 0.05$). This indicates that,

any principled reason, the idiosyncrasies of how this particular speaker realizes Tone 2 immediately after a falling tone.

[6] This was also what most participants reported in their exit survey. They were surprised to learn that the recordings had been edited.

setting the biases aside, the tones are mostly distinct enough from each other. Specifically, the distance between Tone 1 and Tone 2 $D_{T1T2} \approx 0.32$, between Tone 1 and Tone 3 $D_{T1T3} \approx 0.60$, between Tone 2 and Tone 3 $D_{T2T3} \approx 0.73$, between Tone 2 and Tone 4 $D_{T2T4} \approx 0.69$, between Tone 3 and Tone 4 $D_{T3T4} \approx 0.50$. Tone 2 and Tone 3 are the most distinct from each other, while Tone 1 and Tone 2 are the least distinct. Participants can still distinguish tones to a certain extent, even without duration cues. This should not too surprising since other phonetic cues such as amplitude contour are still available in the signals.

The interpretations of the model outputs become somewhat messier from this point onward. Recall that we predicted a stronger bias towards Tone 3 in the LONG-DURATION condition compared to the SHORT-DURATION condition. Since we did not find a significant main effect of duration length, the prediction is not borne out.

However, contrary to expectations, we found significant main effects of **speaker** and **carrier phrase**. The former is easier to understand. Upon closer inspections of the recordings made by Speaker A versus Speaker B, we noticed several systematic differences in duration patterns of the target items. Namely, compared to Speaker B, Speaker A's whispered target words are on average longer. Speaker A's whispered Tone 3 stimuli specifically exhibit more variations in duration. In addition, comparing Speaker A's whispered stimuli to their phonated counterparts, we observe an enhancement of the duration contrasts among the four tones. The two speakers clearly utilized different strategies when producing the whispered tokens, which explains the main effect of the speaker.

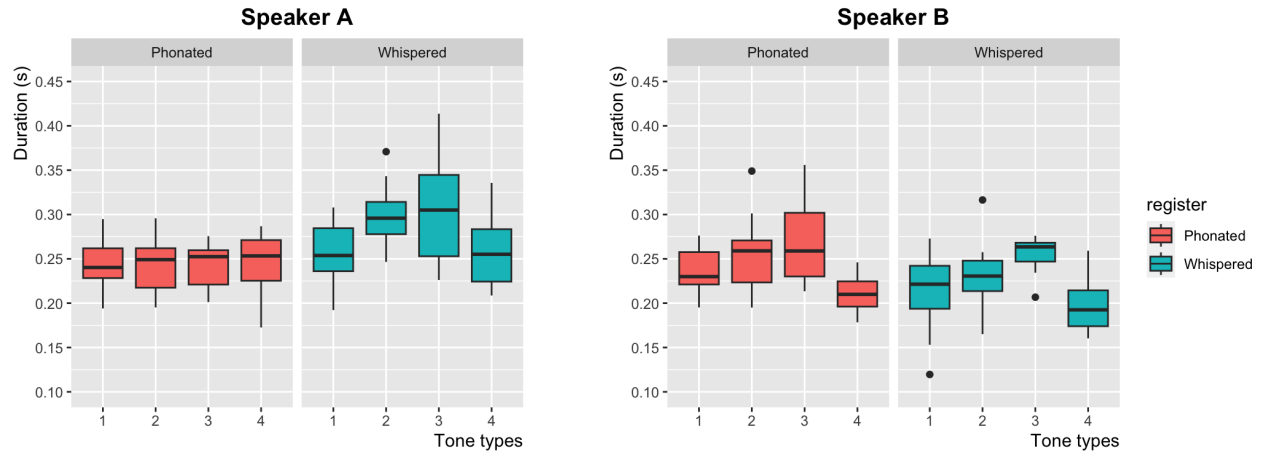

Figure 5. Duration of target items by Speaker A versus Speaker B

The variation between the two carrier phrases is more puzzling. We did not anticipate a co-articulation effect in whispered speech due to its lack of fundamental frequency, but this result seems to suggest otherwise. Since we only included two forms of carrier phrases in this study, we do not know if this is a general pattern and if the effect is robust. It is worth exploring in more controlled experiments in the future.

Finally, we found significant **interactions between responses and carrier phrases**, **between responses and speakers**, and **between responses and duration lengths**. These interactions are sometimes difficult to interpret due to limited variations within each category, but they can still provide some insights into how participants respond to different carrier phrases, speakers, and duration conditions. For instance, participants show a stronger bias towards Tone 3 when

responding to the stimuli recorded by Speaker A. In addition, participants exhibit a weaker bias towards Tone 4 when responding to the LONG-DURATION stimuli. Considering that Speaker A's whispered target words tend to be longer and that Tone 4 typically has a shorter duration, both observations are consistent with our second hypothesis.

Since the data are collected from forty different participants, ideally, we should run a mixed-effects model, adding *participant* as a random effect to account for individual variation. However, the mixed-effects model turns out to be too complicated and fails to converge.

6. Conclusions.

6.1. KEY FINDINGS. We set out to investigate whether, in the absence of the primary cue $F0$, secondary cues provide adequate information for perceiving lexical tones. A few conclusions emerged from our experiment:

First of all, we confirmed that perceptual biases exist and have a substantial effect on accuracy rates. When duration cues are removed, listeners exhibit a strong tendency to categorize a whispered token as either Tone 3 or Tone 2. The origin of this bias remains unclear, as there are no apparent effects of token or type frequency. But what we know is that it is important to look beyond above-the-chance accuracy rates for more robust evidence of listeners' ability to discern phonetic contrasts.

Secondly, we observed that when duration cues are eliminated from the signal, overall performance in tonal identification drops. We arrived at this conclusion through an indirect comparison with the results from similar experiments in Jiao et al. (2015). This suggests that listeners indeed rely on duration cues, at least to a certain extent, for identifying tones in whispered speech.

Thirdly, we found that even after accounting for biases, pairwise distance measures between tones remain significant. This finding implies that lexical tones are still distinguishable to some degree even in the absence of both fundamental frequency and duration cues.

Last but not least, it is still highly debated whether the ability to distinguish tones in whispers is primarily due to speakers intentionally enhancing secondary cues, or the result of listeners becoming more sensitive to said cues in the absence of $F0$. The former is known as the *production enhancement* hypothesis, supported by studies including Liu & Samuel (2004) and Zhang et al. (2022), while the latter is the *perceptual compensation* hypothesis, advocated by Jiao & Xu (2019), among others. We observed that Speaker A exaggerates duration contrasts more in whispered utterances than in phonated speech, which is in line with the production enhancement hypothesis. However, our evidence is extremely preliminary, and further investigations are needed to shed light on this open question.

Despite some promising leads, our study ultimately proves inconclusive. We did not find direct support for the hypothesis that longer words are more likely to be identified as Tone 3 when whispered. Although significant interaction terms provide suggestive evidence, further refinement of the experimental design is necessary.

6.2. CONFOUNDS IN DESIGN. There are a few confounds in the design and execution of this experiment that we would like to address in the future:

- Conduct a baseline experiment with unedited stimuli. Such an experiment would be beneficial for two reasons: (i) to verify if the above-chance results persist when the target words are embedded in carrier phrases, and (ii) to establish the baseline performance of tonal

identification for this specific dataset. This would then allow a direct comparison between the results obtained in our main study and the actual baseline.

- Synthesize duration in increments and aim for more fine-grained results, instead of treating the duration length condition as a binary variable.
- Request separate recordings of carrier phrases and target words. Instead of asking speakers to record entire sentences, we can digitally edit these clips to insert the target words into the carrier phrases. This approach ensures there are no co-articulation effects and maintains consistency in the carrier phrase parts of the stimuli.
- Normalize the amplitude across all items within speech types. Within each phonation type, we can calculate the average root mean square over the end of the tokens and then scale all recordings so that the amplitude of the end of each token is the same. This way, we can rule out amplitude differences as a possible confound while preserving natural amplitude differences between vowels.
- Add a comprehensive exit survey to systematically record participants' impressions of the study. Several participants volunteered interesting comments, such as their self-rated confidence in their responses and their guesses about the speakers' accents

References

Abramson, Arthur S. 1972. Tonal experiments with whispered Thai. In Albert Valdman (ed.), *Papers in linguistics and phonetics to the memory of Pierre Delattre*, 31–44. Berlin/Boston: De Gruyter Mouton (reprint of 2015 edn.). https://doi.org/10.1515/9783110803877-004.

Abramson, Arthur S. 1978. Static and dynamic acoustic cues in distinctive tones. *Language and Speech* 21(4). 319–325. https://doi.org/10.1177/002383097802100406.

Chang, Charles B. & Yao Yao. 2007. Tone production in whispered Mandarin. *UC Berkeley PhonLab Annual Report* 3(3).

Di Paolo, Marianna & Alice Faber. 1990. Phonation differences and the phonetic content of the tense-lax contrast in Utah English. *Language Variation and Change* 2(2). 155–204.

Estes, W. K. 1997. Some reflections on the role of the choice model in theories of categorization, identification, and learning. In A. A. J. Marley (ed.) *Choice, decision, and measurement: Essays in honor of R. Duncan Luce*, 321–328. Mahwah, NJ: Lawrence Erlbaum.

Flemming, Edward. 2020. 24.967 Topics in experimental phonology. Lecture notes.

Gandour, Jackson T. 1978. The perception of ione. In Victoria A. Fromkin (ed.), *Tone: A linguistic survey*, 41–76. New York: Academic Press.

Jensen, Martin Kloster. 1958. Recognition of word tones in whispered speech. *Word* 14(2–3). 187–196. https://doi.org/10.1080/00437956.1958.11659663.

Jiao, Li, Qiuwu Ma, Ting Wang & Yi Xu. 2015. Perceptual cues of whispered tones: Are they really special? *Interspeech-2015*. 2361–2365.

Jiao, Li & Yi Xu. 2019. Whispered Mandarin has no production-enhanced cues for tone and intonation. *Lingua* 218. 24–37. https://doi.org/10.1016/j.lingua.2018.01.004.

Liu, Siyun & Arthur G. Samuel. 2004. Perception of Mandarin lexical tones. when F0 Information is Neutralized. *Language and Speech* 47(2). 109–138. https://doi.org/10.1177/00238309040470020101.

Luce, R. Duncan. 1963. Detection and recognition. In D. Luce (ed.), *Handbook of mathematical psychology*, 1–103. Hoboken, NJ: John Wiley & Sons.

Miller, John D. 1961. Word tone recognition in Vietnamese whispered speech. *Word* 17(1). 11–15. https://doi.org/10.1080/00437956.1961.11659743.

Wassink, Alicia Beckford. 2006. A geometric representation of spectral and temporal vowel features: Quantification of vowel overlap in three linguistic varieties. *The Journal of the Acoustical Society of America* 119(4). 2334–2350. https://doi.org/10.1121/1.2168414.

Wu, Yaru, Martine Adda-Decker & Lori Lamel. 2020. Mandarin lexical tones: A corpus-based study of word length, syllable position and prosodic position on duration. *Interspeech* 2020. 1908–1912.

Yang, Jing, Yu Zhang, Aijun Li & Li Xu. 2017. On the duration of Mandarin tones. *Interspeech* 2017. 1407–1411.

Yip, Moira. 2002. *Tone*. Cambridge: Cambridge University Press.

Zellou, Georgia, Rebecca Scarborough & Renee Kemp. 2020. Secondary phonetic cues in the production of the nasal short-a system in California English. *Interspeech* 2020. 631–635.

Zhang, Hui, Seth Wiener & Lori L. Holt. 2022. Adjustment of cue weighting in speech by speakers and listeners: Evidence from amplitude and duration modifications of Mandarin Chinese tone. *The Journal of the Acoustical Society of America* 151(2). 992–1005. https://doi.org/10.1121/10.0009378.