Mandarin tone identification in whispers

Experiment grant proposal

Agnes Bi

November 24, 2021

I. Motivation and main research questions

Phonetic contrasts are represented via multiple acoustic dimensions, and signaled by multiple cues simultaneously. In the case of tones, it has been long established that fundamental frequency $F0$, defined as the number of cycles per second on 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 a tonal language can still reliably discriminate between various tones with near-ceiling accuracy (Abramson 1978). In other words, $F0$ is a sufficient cue to tonal identification. This leaves us the question of whether the secondary cues serve any function at all in identification or simply come along for the ride. Researchers have mostly ruled out the null hypothesis that primary cues are necessary conditions to successfully identify contrasts (see e.g. Di Paolo and Faber 1990, Wassink 2006, and Zellou, Scarborough, and Kemp 2020). Instead, secondary cues contribute substantially to the identification process, and listeners seem to be sensitive to a weighted combination of various acoustic dimensions.

In this project, I am interested in how much information listeners can extract from the secondary cues alone. More precisely, when listeners are deprived of a key dimension of cues, how do they make use of the remaining acoustic features?

Since $F0$ cues are overwhelmingly powerful 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 mean that listeners do not attend to these dimensions in tonal identification. Whispered speech provides an ideal context for investigation. In whispers, $F0$ is absent since the periodic voicing is replaced by noise source, and, unlike synthetic stimuli with superimposed average $F0$ contours, they are naturally occurring in daily conversations. Using Mandarin as a case study, I propose two perception experiments to directly test for the effect of duration, and to answer the following questions:

- Assuming the most liberal definition of *success* in tonal identification as above-chance accuracy, can listeners still successfully distinguish tones when $F0$ is absent? (Experiment 1)

- If so, what are the secondary cues recruited by listeners in these inhibited conditions, and how are they used? (partially answered by Experiment 2)
- Are there genuine sensitivities to tonal contrasts in whispers? Or are the data accounted for by biases alone? (Experiment 1 & 2)

II. Pilot study

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 and Yao 2007). The typical patterns of the four tones in citation form are illustrated in the figure below, where each tone is plotted with their average duration proportional to the average duration of Tone 3:

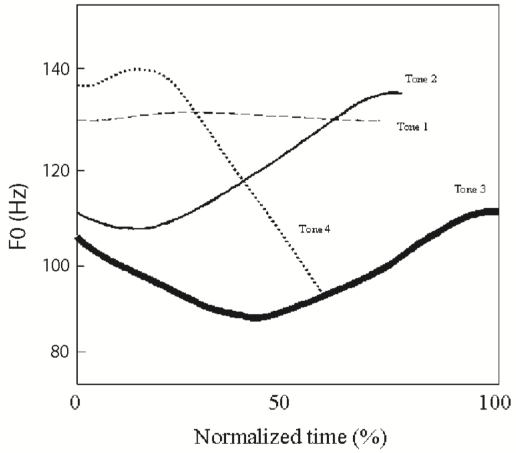

Figure 1: Typical $F0$ contours for the four contrastive tones in Mandarin (Liu and Samuel 2004, adapted from Xu 1997)

Replicating Jiao and Xu 2019 (J&X), I ran a pilot study to test identification of individual words in citation forms. The stimuli are five sets of monosyllables composed of either only vowel ([ɤ] and [u]), or with a glide or nasal onset (/i/-[ji], /y/-/[jy], and /a/-[ma]). These are the same as the ones used in J&X except the /a/ quadruplet. Syllables composed of only /a/ are generally discourse markers with meanings such as 'ah' or 'what?', which correspond to the same character, making it difficult to indicate in a forced-choice task. The five sets of tone quadruplets are shown below:

Tone \ Vowel		/ma/	/ɤ/	/i/	/u/	/y/
T1	Character	妈	婀	衣	乌	迂
	Pinyin	mā	ē	yī	wū	yū
	Glossary	'mother'	'graceful'	'clothes'	'black'	'winding'
T2	Character	麻	鹅	姨	无	鱼
	Pinyin	má	é	yí	wú	yú
	Glossary	'hemp'	'goose'	'aunt'	'nothing'	'fish'
T3	Character	马	恶	椅	五	雨
	Pinyin	mǎ	ě	yǐ	wǔ	yǔ
	Glossary	'horse'	'nausea'	'chair'	'five'	'rain'
T4	Character	骂	饿	意	物	玉
	Pinyin	mà	è	yì	wù	yù
	Glossary	'scold'	'hungry'	'meaning'	'thing'	'jade'

Table 1: Target Stimuli

Four native Mandarin speakers participated in the perception experiment. At each trial, after a stimulus is played, participants are given four options, each of which is a simplified Chinese character corresponding to one of the relevant tone quadruplets. Participants are told to choose the closest character to what they heard.

The results are mostly consistent with what was reported in J&X, and are summed up in the following confusion matrix.

Target \ Response	1	2	3	4
T1	16	7	12	19
T2	8	15	32	7
T3	1	2	71	1
T4	10	4	11	40

Target \ Response	1	2	3	4
T1	29.6%	13.0%	22.2%	35.2%
T2	12.9%	24.2%	51.6%	11.3%
T3	1.3%	2.7%	94.7%	1.3%
T4	15.4%	6.2%	16.9%	61.5%

Table 2: Aggregated confusion matrix in whispered speech

We ran a Biased Choice Model, and found that there is an overwhelming bias towards Tone 3 (bias parameter $b = 0.6709$), and also a smaller yet still substantial one towards T4 ($b = 0.1715$), compared to T1 ($b = 0.0814$) and T2 ($b = 0.0762$). Setting the biases aside, the distances between each tone pairs are still statistically significant, suggesting that speakers are indeed capable of discriminating tones in whispers to certain extent. This begs the questions of how the listeners achieved this and what cues they are relying on. A likely candidate, at least for Mandarin[1], is duration. Unfortunately, I was not able to test this hypothesis with the pilot data – the quiet nature of whispered words makes it difficult to find a consistent criterion to mark their boundaries. Moreover, instead of drawing conclusions from post hoc statistical analyses, which is the common

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

practice in this literature, I will explicitly test the effect of duration by incorporating it into the experimental design.

III. Experiment design

Given the limitations of the pilot study discussed above, we redesign the experiments to (i) embed the target words in carrier phrases; (ii) synthetically manipulate durations of the target words.

The same 20 items listed in Table 1 are embedded in the two carrier phrases (1-a) and (1-b). In (1-a), the word preceding the target item, *shuō* 'say', ends at the upper half of the tonal space, while in (1-b), the offset of the preceding word *chóngfù* 'repeat' ends at the lower half of the tonal space. Even though we do not expect the immediately preceding environment to make a difference for tone identification in whispered speech, it is controlled in case any coarticulation effect does occur.

(1) a. qǐng shuō ___ gěi wǒ tīng.
please say ___ give me listen
'Please say ___ to me.'

b. qǐng chóngfù ___ gěi wǒ tīng.
please repeat ___ give me listen
'Please repeat ___ to me.'

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

The entire sentences, including the carrier portion, will be played to the participants in order to (i) give them a better idea of what the usual speech rate of the speaker is; and (ii) provides some clues of where the boundaries for the target items are.

III.1. Experiment 1: Testing natural stimuli

The main purposes of the first experiment is to establish a baseline performance of tonal identification, and to test whether the above-chance results still hold when the target items are embedded in carrier phrases. It has a full design – each participant will hear and judge all versions of the stimuli listed in 1. The 4 (tones) × 5 (syllables) × 2 (carrier phrases) × 2 (registers) × 2 (speakers) = 160 tokens are completely randomized and divide into two sections to allow for a 5-min break in between. We aim to recruit 20 participants for Experiment 1.

Identification accuracy rate for the phonated tokens is expected to be nearly perfect. We will need to analyze whispered data separately in order to fit a Biased Choice Model. Intuitively speaking, a BCM tells us, when biases are factored out, how well listeners can distinguish tones in whispered register. The dependent variable is the number of responses of each cell in the confusion matrix, and the main predictors are tone types, response categories, and pairwise distance parameters

d. Distance in BCM is essentially a restricted and more interpretable version of `tone:response` interaction in a multinomial logistic regression model.

Fixed effects	tones, carrier phrases, registers
Random effects	syllables, speakers
BCM: count ~ (tone + response + *d*) * (carrier phrase + speaker)	

Assuming the results are consistent with what we found in the pilot and the *d*-values are indeed significant, we will then look closer into how the secondary cues affect tone identification, with a special focus on duration.

III.2. Experiment 2: Testing synthetic stimuli with normalized duration

Experiment 2 takes a novel approach and controls the amount of information listeners receive by holding the duration of the target items constant[2]. This effectively takes away the duration cue, and should lead to worse recognition, i.e., lower accuracy rates, across the board if listeners are indeed heavily relying on duration to discriminate between tones.

Furthermore, we hypothesize that the longer a token is, the more likely it is for a speaker to identify it as Tone 3, which is a directional claim about how duration affects biases in tonal identification. To test this, we construct two levels of the duration variable: (1) the long condition, where durations of all the target item are normalized to the average duration of Tone 4 in their respective register; (2) the short condition, where they are normalized to the average duration of Tone 3. Recall that the average duration of Tone 4 is the longest naturally occurring one, while the average duration of Tone 3 is the shortest.

This gives us 4 (tones) × 5 (syllables) × 2 (carrier phrases) × 2 (registers) × 2 (speakers) × 2 (duration levels) = 320 tokens in total, which is too many to fit into one session, especially considering the repetitive nature of the study. Hence, we decide to make speakers (volunteers who recorded the original stimuli) a between-subject condition, and divide Experiment 2 into two groups: speaker 1 group vs. speaker 2 group. The purpose of the speaker variable is to rule out the possibility that whatever the results we get are due to idiosyncrasies of the particular speaker instead of a general pattern of the language. We aim to recruit 20 participants per group as well.

Experiment Design	
Between subject conditions	speakers
Within subject conditions	tones, syllables, carrier phrases, registers, long vs. short
Statistic Analysis	
Fixed effects	tones, carrier phrases, registers, long vs. short
Random effects	syllables, speakers
BCM: count ~ (original tone + response + *d*) * (duration + carrier phrase + speaker)	

[2]Given the time constraint of the generals requirement, I will collect data for Experiment 2 first since it is the more theoretically interesting one.

Note that the main questions asked for Experiment 1 versus Experiment 2 are largely independent – while Experiment 1 aims to find out how well people can discriminate whispered tones, Experiment 2 is more interested in the inner workings of how specific cues influence biases. However, since the same set of stimuli are used for both, we hope eventually to be able to make at least qualitative comparisons between the two experiments.

IV. Itemized budget and budget justification

The experiments will be conducted at the Behavioral Research Lab facility with the help of ESSL. I intend to recruit $20 + 2 * 20 = 60$ subjects in total. Since each experiment will take approximately 30 minutes to complete, I will pay participants the standard rate of \$7 per person in cash or with gift cards. To further incentivize people to come into the lab, I will enter all participants into a lottery for three \$50 Amazon gift cards. The total budget is thus $60 \times \$7 + \$150 = \$570$.

References

Abramson, A. S. (1978). Static and dynamic acoustic cues in distinctive tones. *Language and Speech*, *21*(4), 319–325.

Chang, C. B., & Yao, Y. (2007). Tone Production in Whispered Mandarin. *UC Berkeley PhonLab Annual Report*, *3*(3).

Di Paolo, M., & Faber, A. (1990). Phonation differences and the phonetic content of the tense-lax contrast in Utah English. *Language variation and change*, *2*(2), 155–204.

Gandour, J. T. (1978). The Perception of Tone. In V. A. Fromkin (Ed.), *Tone: A Linguistic Survey* (pp. 41–76). New York: Academic Press.

Jiao, L., & Xu, Y. (2019). Whispered Mandarin has no production-enhanced cues for tone and intonation. *Lingua*, *218*, 24–37.

Liu, S., & Samuel, A. G. (2004). Perception of Mandarin Lexical Tones when F0 Information is Neutralized. *Language and Speech*, *47*(2), 109–138.

Wassink, A. B. (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.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, *25*(1), 61–83.

Yip, M. (2002). *Tone*. Cambridge University Press.

Zellou, G., Scarborough, R., & Kemp, R. (2020). Secondary Phonetic Cues in the Production of the Nasal Short-a System in California English. In *INTERSPEECH* (pp. 631–635).