How secondary cues contribute to tone identification

Agnes Bi

October 10, 2021

1 Motivation & Questions

It has long been established that fundamental frequency $F0$ serves as the primary cue to tone recognition (Gandour 1978, Yip 2002). 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 (Fok 1974, Abramson 1978). A natural question to ask is then if the primary cue is taken away instead, can speakers differentiate tones with reasonable accuracy? In other words, is F_0 a necessary condition for tone classification?

Whispered speech provides an ideal context for investigation. In whispers, fundamental frequency is absent since the periodic voicing is replaced by noise source, and, unlike synthetic stimuli, they are naturally occurring in daily conversations. This paper presents results from a pilot perception experiment, using Mandarin as a case study. The specific research questions of interest are:

1. What is the overall accurate rate of tonal discrimination in whispers?
2. If it is significantly above chance, what are the secondary cues contributing to this process?
3. Do speakers show biases towards a certain tone?

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 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 proporation to the average duration of Tone 3 (Liu and Samuel 2004, adapted from Xu 1997):

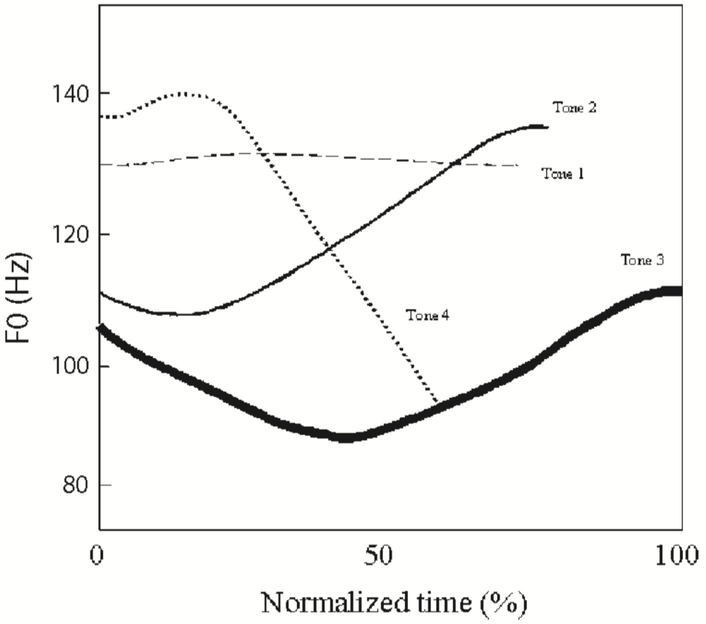

Figure 1: Typical $F0$ contours for the four contrastive tones in Mandarin

2.2 Previous studies

Previous perception studies show a range of accuracy rates of tone perception in whispered speech: some studies have reported performance slightly higher than chance (Abramson 1972 on Thai; Miller 1961 on Vietnamese), and in a few cases, well above chance (Jensen 1958 on Norwegian, Swedish, Slovenian, and Mandarin). There are two potential reasons for this discrepancy in the literature: (1) early studies have varying scope of data and often do not examine the full paradigms; (2) for the statistical analysis part, only *percent correct* is used for evaluating whether the identification is successful, which fails to isolate potential bias.

Liu and Samuel 2004 is one of the first comprehensive experiments

On a related note, neural network models trained without explicit F_0 information manually encoded, such as the ones tested in Ryant, Slaney, Liberman, Shriberg, and Yuan 2014 and in Chen, Bunescu, Xu, and Liu 2016, perform remarkably well in tone classification for phonated speech. This may suggest that some non-F_0 phonetic dimensions are jointly utilized by the models to predict tonal category.

The main goal of this project is to directly probe the two secondary cues, namely *duration* and *temporal envelope*, that have been shown to vary with lexical tones (Tseng 1981, Fu and Zeng 2000, Kong and Zeng 2006, a.o.).

3 Methods

The present study is a slightly simplified replicate of a perception experiments presented in Jiao and Xu 2019 (J&X). Unlike J&X, we did not test intonation as a dimension of variable.

3.1 Stimuli

The perception stimuli are five sets of monosyllables composed of only vowels (/ɤ/ and /u/), with glide onsets (/i/-[ji] and /y/-/[jy]) or with nasal onset (/a/-ma). They are the same stimuli as the ones used in J&X except the /a/ quadruplet, since syllables composed of only /a/ are generally discourse markers with meanings such as 'ah' or 'what?' and they correspond to the same character, which makes it difficult to indicate in the forced-choice task. The five sets of tone quadruplets are shown in Table 1.

Vowel Tone		/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

The stimuli are recorded by two female native speakers of Mandarin, who grew up in eastern coastal China. Speakers read through the same word list in different randomized order. For each speaker, 5 (syllables) * 4 (tones) * 2 (speech types) * 2 (repetition) = 80 stimuli are recorded. Speakers are instructed to read through the list alternating between normal speech and whisper for each token. The phonated version is read first in the first repetition, and whispered first in the second repetition.

The recordings are made in a sound-proof booth at a sampling rate of 44.1kHz as 16-bit format mono sound files.

3.2 Participants

Four native Mandarin speakers between the age of 22-27 participated in the perception experiment, among which two are females and two are males.[1] None of the participants reported any hearing or speech impairment.

[1] Data from another female speaker was collected but not included in the subsequent statistical analysis, as her dominant language is Cantonese, instead of Mandarin.

3.3 Procedure

The written instructions were given in English, but explained again in Mandarin prior to the practice trials. After a stimulus is played, participants are given four options, each is a simplified Chinese character corresponding to one of the relevant tone quadruplets. They were told to choose the closest character to what they heard. Each audio can be repeated at most once. This is to avoid the situation when the participant got distracted and accidentally missed the word. Participants are encouraged to take a break halfway through the experiment. The entire study lasts approximately 15 minutes.

4 Analyses

4.1 Data pruning

A total of 596 observations, excluding practice trials, are collected. The reaction time data in seconds look as follows:

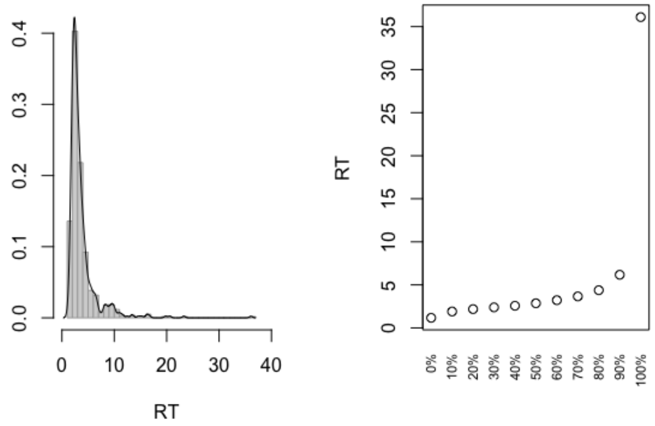

Figure 2: Distribution of reaction time data

95% of the data points fall between 1.17s and 9.11s. Since there are some extreme outliers, the observations with RT greater than 10s are discarded, which is about 3.5% of the data.

4.2 Accuracy & confusion matrix

As expected, tones are generally much worse identified in whispered than in phonated utterances (55.5% vs. 99.4%). In phonated utterances, all tones were perceived essentially at ceiling. In whispers, the picture is more complicated. The identification rate drops to almost chance level for T1 (29.6%) and T2 (24.2%), but much less for T4 (61.5%), which is consistent with the findings in J&X (23.86%, 31.06%, and 60.23%, respectively). However, at least on the surface, T3 identification is, surprisingly, still near ceiling at 94.7% (cf. 84.47% in J&X).

Perceived Produced	1	2	3	4
T1	16	7	12	19
T2	8	15	32	7
T3	1	2	71	1
T4	10	4	11	40

Perceived Produced	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

Perceived Produced	1	2	3	4
T1	10	4	5	9
T2	3	12	10	4
T3	1	2	34	1
T4	2	2	6	24

Perceived Produced	1	2	3	4
T1	35.7%	14.3%	17.9%	32.1%
T2	10.3%	41.4%	34.5%	13.8%
T3	2.6%	5.3%	89.5%	2.6%
T4	5.9%	5.9%	17.6%	70.6%

Table 3: Confusion matrix in whispered speech of female participants

The pattern of confusion in tone recognition is quite interesting in itself, and can potentially provide some insights into how the four tones occupy the abstract tonal space. When Tone 1 stimuli were presented, listeners more often identified it as Tone 4 than itself. Strikingly, Tone 2 is identified much more often as Tone 3 than itself, but Tone 3 is rarely misidentified as Tone 2. The overwhelming bias toward Tone 3 is not observed in either Liu and Samuel 2004 or Jiao and Xu 2019. In the next subsection, I will venture a guess to the cause of this disparity.

4.2.1 Possibly, partition of speakers?

I would like to state upfront that what the results in this section might suggest cannot be taken at face value considering the small sample size.

By fitting a binomial logistic regression model with accuracy rate as dependent variable and speech type, tone, participant gender, stimulus duration as independent variables (C = 0.942), we find that participant gender is a statistically significant predictor of accuracy rate (with gender = female being the reference level, b = -0.7168, SE = 0.3085, $p = 0.0201 < 0.05$), which seems quite puzzling. I hypothesize that there are two distinct classes of speakers in terms of how they deal with “confusing” tones in whispers, Tone 1 and Tone 2 in particular, and it just happens to be the case that in this particular experiment, the two natural classes coincide with gender divide. Class I listeners are equally likely to judge Tone 1 as Tone 1 or Tone 4, and to judge Tone 2 as Tone 2 or Tone 4. The pattern would be similar to what’s shown in Table 5:

Class II listeners are more inclined to default to Tone 3 or Tone 4, when presented with Tone 2 or Tone 1 stimuli in whispers. The pattern in Table 6 falls along this line:

Further studies drawn from a much larger sample size are needed to test this hypothesis. If it does have some empirical ground, we could argue for some intrinsic connection between Tone 1 and Tone 4, and between Tone 2 and Tone 3 that make them pair-wise more easily confusable.

Perceived Produced	1	2	3	4
T1	6	3	7	10
T2	5	3	22	3
T3	0	0	37	0
T4	8	2	5	16

Perceived Produced	1	2	3	4
T1	23.1%	11.5%	26.9%	38.5%
T2	15.2%	9.1%	66.7%	9.1%
T3	0%	0%	100%	0%
T4	25.8%	6.5%	16.1%	51.6%

Table 4: Confusion matrix in whispered speech of male participants

4.3 Perceptual bias and distances

As expected, tone identification is much less accurate in whispered than in phonated utterances (55.5% vs. 99.4%). In phonated utterances, all tones were perceived essentially at ceiling. In whispers, the picture is more complicated. The identification rate drops to almost chance level for T1 (29.6%) and T2 (24.2%), but less so for T4 (61.5%), which is consistent with the findings in J&X (23.86%, 31.06%, and 60.23%, respectively). However, at least on the surface, T3 identification is, surprisingly, still near ceiling at 94.7% (cf. 84.47% in J&X).

However, the high accuracy rate in identifying T3 should not be taken at face value, since it is likely due to a bias defaulting to T3 when the participant is uncertain. A Biased Choice loglinear model, with the baseline for response set at T3, is fitted to test this hypothesis.

From the output above, we can estimate bias parameters for each tone, and calculate estimated perceptual distance between each pair of tones.

Let's first calculate the bias parameters. Since all three response coefficients have $p \ll 0.05$, we can safely conclude that the bias towards T1, T2, and T4 are significantly different from that towards T3. Normalizing the e^{λ} values gives us the BCM bias parameters.

	loglinear coefficient λ	e^{λ}	bias parameter b
T1	-2.109372	0.12131412778	0.0814
T2	-2.175254	0.11357930086	0.0762
T3	0	1	0.6709
T4	-1.364372	0.25554110542	0.1715

Table 5: Bias parameters calculation

There is an overwhelming bias towards T3, which was not noted in either Liu and Samuel 2004 or Jiao and Xu 2019.

Although d values obtained in BCM are not quite “distances" in the usual sense, it is interesting to note that even in whispered speech, each pair of tones are still distinct enough (all 6 pairs with $p < 0.05$) from each other. Specifically, $D_{\text{T1T2}} = 0.725981$, $D_{\text{T1T3}} = 2.380275$, $D_{\text{T1T4}} = 0.597176$, $D_{\text{T2T3}} = 1.416213$, $D_{\text{T2T4}} = 1.544211$, $D_{\text{T3T4}} = 2.678014$, with T3 and T4 being most distinct from each other and T1 and T4 the least.

A closer look into each participant's data seems to suggest a difference in bias. Hence, participant is added as an additional variable in BCM to test the effects. Unfortunately, given the limited data, 3 coefficients are not defined because of singularities.

In sum, BCM reveals a strong bias towards T3, but tones in whispered speech are still perceptually distinct. This encourages further investigation into what, then, are the cues speakers rely on for tone identification when F_0 is taken away. The two obvious candidates are duration and amplitude, and I will explore the former in the next subsection.

4.4 Reaction Time

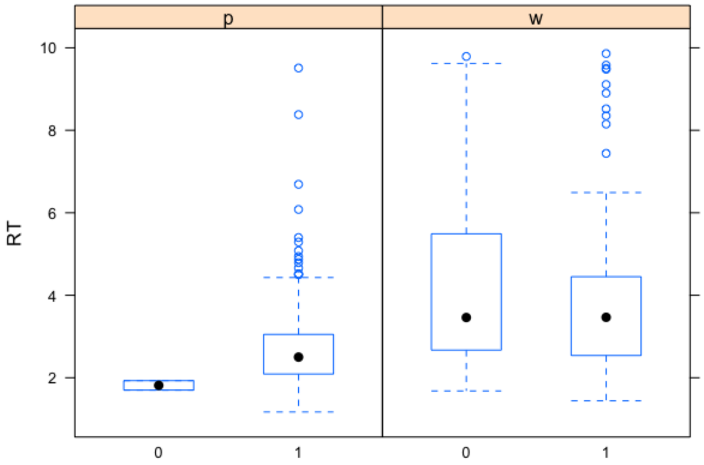

Figure 3: Reaction time distribution given accuracy of the response

It does not come as a surprise that in general, the reaction time for whispered speech is significant longer than the RT for phonated speech. It is worth noting that based on the short RT for the two misidentified phonated tokens, the mistakes are most likely due to technical error such as pressing the wrong key than genuine confusion.

4.5 Duration

Many studies have suggested temporal envelope and intensity being possible cues for tone recognition in whispers. Liu and Samuel 2004 find a correlation between syllable duration and tone perception. This subsection presents some preliminary results as I try to look into the correlation between accuracy rate and duration of the stimuli.

The duration data used here are audio clip length extracted automatically using a python script. The recordings were not quite segmented based on a consistent criterion, since the beginning and ending of a whispered sound is oftentimes difficult to tell on the spectrogram. Due to time limitation, I was not able to go back to the original recordings and measure duration of each token manually, but that will be the next step for this project.

A box and whisker plot is given below; 0 means incorrect response, and the y-axis is the duration of the corresponding set of stimuli in seconds.

An ANOVA test is performed to determine whether leaving the variable Duration out of the binomial logistic regression model would significantly reduce the deviance. Duration is not shown to be a significant predictor of accuracy rate ($p = 0.4279 \gg 0.05$).

5 Analysis

5.1 Duration

Many studies have suggested temporal envelope and intensity being possible cues for tone recognition in whispers. Liu and Samuel 2004 find a correlation between syllable duration and tone perception. This subsection presents some preliminary results as I look into the correlation between response and duration[2] of the stimuli.

Let's hypothesize that a longer duration should result in participants favoring T3 and disfavoring T4. A multinomial logistic regression is used to model the decision process, with stimuli and duration as predictors.

I'm not quite sure how to interpret the results of a multinomial logistic regression, but a model with only stimulus type as the predictor seems to give a better fit (lower AIC), which contradicts our intuition.

6 Discussion

To summarize, Tone 3 and Tone 4 overall show significantly above chance identification rates in whispers, with the former near ceiling. Tone 1 and Tone 2 identification are essentially at chance.

In whispered speech, participants show a strong bias towards Tone 3 and a weak one towards Tone 4. However, the four tones seem to still be perceptually distinct from each other, the reason of which remains unclear.

Two interesting perceptual asymmetries are observed: (a) Tone 2 is much more frequently misidentified as Tone 3 than Tone 3 as Tone 2; and (b) Tone 1 is more often misidentified as Tone 4 than Tone 4 as Tone 1. The similarity between Tone 2 and Tone 3 has been argued for extensively in literature (Shen 1989, Liu and Samuel 2004, a.o.), but the connections between Tone 1 and Tone 4 need further investigation. An obvious one is that both Tone 1 (55) and Tone 4 (51) have high pitch register onsets. Since amplitude contour mirrors F0 contour (Whalen and Xu 1992), one could argue that onset amplitude is an important cue to tone identification in whispers, as it first bipartitions logical space to either choosing between Tone 1 and Tone 4 if the onset amplitude is high, or choosing between Tone 2 and Tone 3 if the onset amplitude is low. The confusion between Tone 1 and Tone 4 and between Tone 2 and Tone 3 seem to suggest that listeners are less sensitive to the overall shape of the amplitude contour. Moreover, assuming complex contour tones (Tone 3) are structurally more complicated than simple contour tones (Tone 2, Tone 4), and in turn, more complicated than level tones (Tone 1), when there is insufficient information regarding the overall shape of the amplitude contour or when listeners cannot determine it for certain, at least a subset of the listeners defaults to the more complicated structure (hypothesized Class II speakers in section IV.2.1).

[2] The duration data used here are audio clip length extracted automatically using a python script. The recordings were not quite segmented based on a consistent criterion, since the beginning and ending of a whispered sound is oftentimes difficult to pinpoint on the spectrogram. A better approach would be to measure each token duration manually.

6.1 Potential Confounds

To end this squib, I will point out some potential confounds of the present study and what I have learned from this pilot experiment.

First of all, all the response options given in each trial are characters corresponding to minimal tone quadruplets, which could make research objective quite transparent to the participants. This is a drawback in the experimental design, and in future experiments, I should incorporate response options with, for example, varying syllables as filler items.

Secondly, since the audio clips are recorded separately by two different speakers, the average amplitude within each speech type should be normalized across speakers.[3]

Thirdly, it is nontrivial to establish a consistent segmenting criterion for the whispered tokens, as their left and right boundaries are often fuzzy. Directly comparing acoustic properties of whispered syllables with their phonated counterparts can be tricky. In the present study, the boundaries of a whispered sound are largely determined by its spectrogram shape, and when there are clear formants, they seem to be consistent with the phonated version. Note in many cases, Praat cannot generate reliable formant readings (for f1 in particular), which suggests that automated data extraction is likely to fail.

[3] Even though there was not a main effect of Speaker in the findings here, it is probably in general a good practice to control individual variation.

References

Abramson, A. S. (1972). Tonal Experiments with Whispered Thai. In *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre* (Reprint 2015, pp. 31–44).

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).

Chen, C., Bunescu, R. C., Xu, L., & Liu, C. (2016). Tone Classification in Mandarin Chinese Using Convolutional Neural Networks. In *Interspeech* (pp. 2150–2154).

Fok, C. Y.-y. (1974). *A perceptual study of tones in Cantonese*. Hong Kong: Centre of Asian Studies, University of Hong Kong.

Fu, Q.-J., & Zeng, F.-G. (2000). Identification of temporal envelope cues in Chinese tone recognition. *Asia Pacific Journal of Speech, Language and Hearing*, *5*(1), 45–57.

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

Jensen, M. K. (1958). Recognition of Word Tones in Whispered Speech. *WORD*, *14*(2-3), 187–196.

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

Kong, Y.-Y., & Zeng, F.-G. (2006). Temporal and spectral cues in Mandarin tone recognition. *The Journal of the Acoustical Society of America*, *120*(5), 2830–2840.

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

Miller, J. D. (1961). Word Tone Recognition in Vietnamese Whispered Speech. *WORD*, *17*(1), 11–15.

Ryant, N., Slaney, M., Liberman, M., Shriberg, E., & Yuan, J. (2014). Highly accurate mandarin tone classification in the absence of pitch information. In *Proceedings of Speech Prosody* (Vol. 7), Citeseer.

Shen, X.-n. S. (1989). *The Prosody of Mandarin Chinese*. Berkeley: University of California Press.

Tseng, C.-Y. (1981). *An Acoustic Phonetic Study on Tones in Mandarin Chinese* (Doctoral dissertation, Brown University, United States – Rhode Island).

Whalen, D. H., & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica*, *49*(1), 25–47.

Xu, Y. (1997). Contextual tonal variations in Mandarin.

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