Perceptual cues of whispered tones: Are they really special?

Li Jiao[1], *Qiuwu Ma*[1], *Ting Wang*[1], *Yi Xu*[2]

[1] School of Foreign Languages, Tongji University, Shanghai, China
[2] Department of Speech, Hearing and Phonetic Sciences, University College London, London, UK
{2013li_jiao; mqw; 2011ting_wang}@tongji.edu.cn, yi.xu@ucl.ac.uk

Abstract

This paper reports our initial findings on whether Mandarin Chinese has developed effective strategies to convey tonal information in whispered speech. We recorded phonated and whispered tones in monosyllabic words, and analyzed the acoustic properties of the tonal contrasts. We then generated amplitude-modulated noise based on both the phonated and whispered utterances and used them as stimuli in a tone perception experiment, together with the original phonated and whispered speech. Results showed that, once turned into amplitude-modulated noise, originally phonated and whispered speech had similar perceptual patterns, and that these patterns resembled those of whispered tones. The acoustic analysis showed that properties corresponding to tonal contrasts in whispered speech already existed in the phonated tones, and there was no evidence of enhancement of these properties in whispering. Overall, therefore, Mandarin may not have developed highly special strategies to enhance tonal contrast in whispered speech.

Index Terms: Tone perception, whispered speech, phonated speech, amplitude-modulated noise

1. Introduction

Given the importance of lexical tone in a language like Mandarin, it is natural to assume that speakers of the language have developed special strategies to enhance the tonal contrast in whispered speech as there is no voicing to carry F_0, the crucial acoustic correlate of tone. Much research has been conducted in search of properties that enable the perception of tone and intonation in whispering, and many of them have been reported. These include duration [1, 2], intensity [3], formants [3, 4, 5] and spectral tilt [6].

A question that has not been closely examined, however, is whether the reported tone-cuing properties are special to the whisper register, or they also exist in phonated speech. There has already been some evidence for the latter possibility. It has been found, for example, that both duration and amplitude profiles provide tonal information when F_0 is absent in signal-correlated noise [7], but their contribution to tone perception is small (about 3%) when F_0 is present [8].

Another reason for questioning the special-cue account of whispered tone is that, although the functional load of tones is about the same as vowels in Mandarin [9], the intelligibility of news-like speech remains unaffected when F_0 is flattened, and it is only when signal-to-noise ratio is significantly reduced that intelligibility starts to deteriorate [10]. That finding suggests that cue redundancy in speech is high so that an entire dimension, such as tone, can be absent without affecting communication in a relatively ideal listening environment. Such redundancy is needed in adverse conditions such as at cocktail parties [11].

The present study is a preliminary attempt to examine if Mandarin has developed or at least enhances secondary cues to lexical tones. This was done by making direct acoustic comparisons of phonated and whispered tones. We also converted both phonated and whispered utterances to amplitude-modulated noise and examined their tonal intelligibility. Overall, our goal was to answer the following research questions:

1. Are Mandarin tones still intelligible in whispered speech?
2. If yes, are there identifiable acoustic cues for the perception of the whispered tones?
3. If yes, are these cues special to whispered tones, or are they already present in the phonated tones?

2. Methods

2.1. Speaking materials

Table 1 is a list of all the target syllables in the database used in the present study. The database consists of five sets of syllables with vowel or glide onsets and three sets with obstruent onsets. They carry the four lexical tones of Mandarin: Tone 1 (T1), Tone 2 (T2), Tone 3 (T3) and Tone 4 (T4). All syllables of the database were subjected to acoustic analysis in the present study, but only a subset, i.e., those in columns 2, 3 and 5, were used in the perception experiment. Those syllables are either vowel-only: e [ɤ], or with glides as onsets: yi, yü. These three syllables were chosen because a) they would lead to the least undesirable artifact in generating amplitude-modulated noise, b) there were no distinct characters to represent /a/ with four different tones, which makes the perception test difficult, and c) the signal to noise ratio of whispered /u/ was not high enough to prevent unnecessary confusion.

Table 1. *A list of syllables in the database.*

Vowel Tone		a	ɤ	i	u	y	a	ɿ	ʅ
T1	Character	啊	婀	衣	乌	迂	插	疵	吃
	Pinyin	*ā*	*ē*	*yī*	*wū*	*yū*	*chā*	*cī*	*chī*
	Glossary	*oh*	*graceful*	*clothes*	*black*	*winding*	*insert*	*flaw*	*eat*
T2	Character	啊	鹅	姨	无	鱼	茶	词	迟
	Pinyin	*á*	*é*	*yí*	*wú*	*yú*	*chá*	*cí*	*chí*
	Glossary	*eh*	*goose*	*aunt*	*nothing*	*fish*	*tea*	*word*	*late*
T3	Character	啊	恶	椅	五	雨	衩	此	尺
	Pinyin	*ǎ*	*ě*	*yǐ*	*wǔ*	*yǔ*	*chǎ*	*cǐ*	*chǐ*
	Glossary	*what*	*nausea*	*chair*	*five*	*rain*	*underpants*	*this*	*ruler*
T4	Character	啊	饿	意	物	玉	岔	次	赤
	Pinyin	*à*	*è*	*yì*	*wù*	*yù*	*chà*	*cì*	*chì*
	Glossary	*ah*	*hungry*	*meaning*	*thing*	*jade*	*branch*	*sequence*	*red*

2.2. Recording Procedures

These syllables were recorded either with or without a carrier (‘*zhè gè zì niàn …*’ ‘*The word is read as…*’), and in three modes: citation, interactive statement, and interactive question. In the citation session, the speakers were asked to simply read aloud the Chinese characters one by one. In the interactive sessions, the speakers were asked to “perform,” with one saying the key word as a question, and the other saying the same word as an answer. Examples are given below. The speakers rotated their question-answer roles between trials. Before each session, there was a short round of practice.

Speaker 1: é ?	*“goose?”*	(question)
Speaker 2: é.	*“goose.”*	(statement)
Speaker 2: yǐ ?	*“chair?”*	(question)
Speaker 1: yǐ .	*“chair.”*	(statement)

A total of 8 syllables * 4 tones * 2 registers (phonated / whisper) * 3 modes (citation / statement / question) * 2 carrier conditions * 2 repetitions = 768 target syllables were recorded by each speaker.

Two native speakers of Mandarin (a male and a female, mean age = 26.5 years) took part in the recording. They sat side by side in the sound booth in the Phonetics Laboratory, University of Oxford. The recording was done with an Audio-Technica AT4031 microphone, which was on a stand between the two speakers, with a distance of 15 cm from each. A software tool was used to present the stimuli (characters and corresponding pinyin) on a screen inside the sound booth. The experimenter monitored the recording and controlled the progression of the recording outside the booth. The input volume of the recording was set to be the same for phonated and whispered registers, and we made sure that it was neither too loud for the phonated register nor too soft for the whispered register. The sounds were recorded onto a Compact Disk by a CD recorder (HHB CDR-850) at 44.1 Hz and 16 bits resolution, and then re-recorded into a PC using a Sound Blaster analogue to digital conversion.

2.3. Perception stimuli

The phonated and whispered syllables e, yi, and yü by the female speaker were used as stimuli for the perception experiment. For the natural speech condition, the original recordings were used. For the amplitude-modulated noise condition, the AmplitudeTier of each syllable was extracted in Praat [12] and imposed onto pink noise of the same duration. The pink noise was generated by filtering white noise in Praat with a -6 dB/octave de-emphasis starting from 50 Hz. The resulting signal therefore contained both the duration and amplitude undulation of the syllable, but no spectral information. The amplitude-modulated noise was generated for both the phonated and whispered syllables. During the process, the maximum absolute amplitude was scaled to 0.9, which neutralized the difference between the phonated and whispered tone in overall intensity.

2.4. Perception procedure

Twenty native Mandarin listeners (10 males and 10 females) living in China participated as subjects. They were aged from 18 to 27 (with a mean of 20.3 years), and had no self reported speech and hearing disorders. The experiments took place in a quiet room in Tongji University, Shanghai. Subjects wore Sennheiser PC166 headphones, and were seated comfortably in front of a Dell computer (OPTIPLEX 390). At the beginning of each session, they received instructions and had a short round of practice.

The tests were run with an ExperimentFMC script in Praat. In each trial, the subject heard an utterance (or noise), and saw on the screen four Chinese characters of the corresponding syllables with four tones. They then pressed the button with the character closest to what they had heard. Each sound was played only once. The test took around 30 to 40 minutes to complete. All the subjects were tested first with the natural speech stimuli, and then with the amplitude-modulated noise. For the natural speech tasks, half of the subjects heard the phonated syllables first, while the other half heard the whispered syllables first.

In total, each subject went through 288 trials: 3 syllables * 4 tones * 2 registers (phonated / whisper) * 3 modes (citation / statement / question) * 1 carrier condition (without a carrier) * 2 signal types (speech / noise) * 2 repetitions.

3. Results

3.1. Acoustic cues of whispered tones

The acoustic analysis was done with a modified version of ProsodyPro, a Praat script for large-scale prosody analysis [13]. With the script, we annotated the utterances by the female speaker. The script then generated the following measurements:

Duration (ms) — Duration of target syllable

F1, F2, F3 (Hz) — Frequencies of first three formants at syllable center

Intensity (dB) — Mean intensity of target syllable

Temporal profile of intensity (dB) — 10-point time-normalized intensity contour

Spectral center of gravity (COG) — Centre of spectral gravity

Hammarberg index — Difference between the maximum energy in the 0.2kHz and 2.5kHz bands [14]

Energy below 500, 1000 Hz — Energy of voiced segments below 500Hz and 1000Hz

These measurements were analyzed in a set of 3-way ANOVAs, with phonation (phonated, whispered), mode (citation, declarative, interrogative) and tone (1-4) as independent variables. Due to space limit, however, we will not discuss the effect of speaking mode (citation, statement, question).

Figure 1 shows the effects of phonation and tone on duration. In terms of phonation, phonated tones are shorter than whispered tones (F(1,744) = 421.84, $p < 0.0001$). For tone there is a significant main effect (F(3,744) = 188.85, $p < 0.0001$). A Student-Newman-Keuls post-hoc test found all the four tones to be significantly different from each other in duration. However, there is no interaction between phonation and tone, suggesting that, other than being generally longer, the whispered tones did not show duration patterns different from those of phonated tones.

For the formants, no main effect of tone was found. But there were significant effects of phonation. Compared to phonated formants, whispered F1 was higher (F(1,744) = 371.16, $p <$

0.0001), whispered F2 was also higher (F(1,744) = 8.54, p = 0.0036), but whispered F3 was lower (F(1,744) = 29.13, p < 0.0001). Importantly, there were no interactions between phonation and tone for any of the formants.

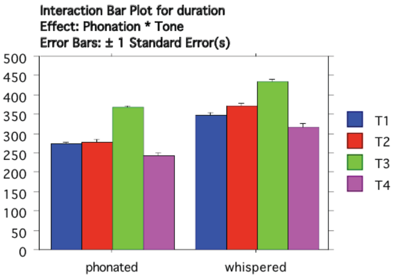

Figure 1: *Means and standard errors of syllable duration in phonated and whispered tones.*

Figure 2 shows the effect of phonation and tone on the mean intensity of the syllables. ANOVA results showed that phonated syllables had significantly higher intensity than whispered syllables (F(1,744) = 4314.08, p < 0.0001). There was also a main effect of tone (F(3,744) = 9.54, p < 0.0001). But a Student-Newman-Keuls post-hoc test found significant difference only between T3 and each of the other three tones. More importantly, there is again no interaction between phonation and tone.

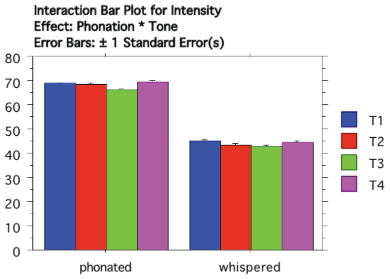

Figure 2: *Means and standard errors of intensity as a function of phonation and tone.*

To see if there are more detailed tone-specific intensity patterns, Figure 3 displays mean time-nomalized temporal profiles of intensity of the four tones in phonated as well as whispered speech. As can be seen, within the same tone, the profiles do not differ much between the two phonation registers. Across the tones, only T3 stands out with a bimodal profile, and T4 with a slightly greater drop in the later half of the syllable.

With regard to spectral tilt, of the four measurements we took, only two showed main effects of tone: Hammarberg index (F(3,744) = 5.24, p = 0.0014) and energy below 500 Hz (F(3,744) = 3.84, p = 0.0096). Again, there was no interaction of phonation and tone, as the tonal differences showed similar pattern in both phonation registers, as can be seen in Figure 4.

Overall, therefore, we have found acoustic properties corresponding to the four tones of Mandarin other than F_0, in terms of duration, intensity and spectral tilt. However, these properties occurred in both phonated and whispered speech, and there do not seem to be cues that are special to whispered tones.

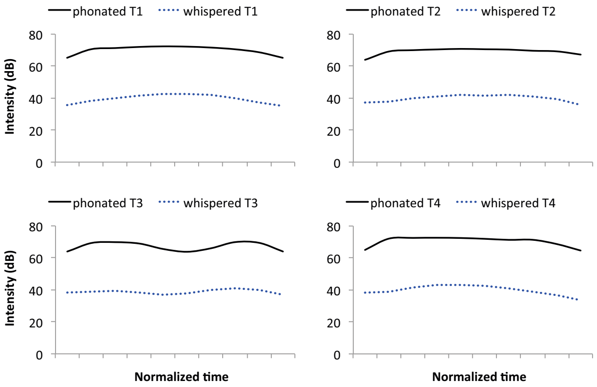

Figure 3: *Mean temporal profiles of intensity of phonated (solid lines) and whispered tones (dotted lines). Each profile is an averaged of 18 tokens of a tone in the given phonation register.*

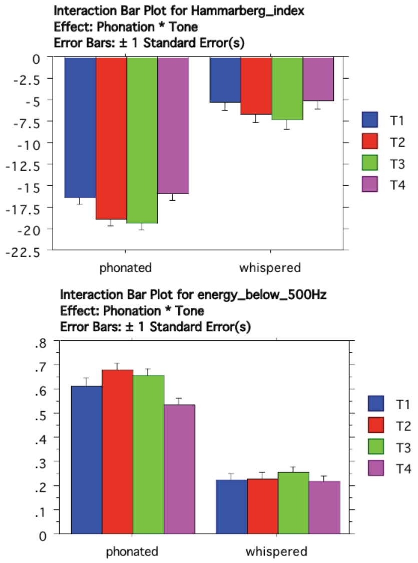

Figure 4: *Means and standard errors of Hammarberg Index (top) and energy below 500 Hz as functions of phonation and tone.*

3.2. Perceptual cues of whispered tones

Figure 5 shows overall identification rates for both phonated and whispered tones from original (upper panel) and amplitude-modulated noise (lower panel). The identification rates of the phonated tones were near or at ceiling across the board. For the original speech utterances, a three-way repeated measures ANOVA showed significant main effects of phonation (F(1,19) = 1576.62, p < 0.0001), tone (F(3,57) = 56.23, p < 0.0001) and modality (F(2,38) = 12.42, p < 0.0001). There was a significant interaction between phonation and tone (F(3,57) = 88.66, p < 0.0001), showing that phonated and whispered tones were perceived rather differently. This interaction can be clearly seen in the top panel of Figure 5. (Again we will not discuss the effect of speaking mode on tone perception in this paper.)

For the amplitude-modulated noise, a similar three-way repeated measures ANOVA showed significant main effect of tone (F(3,57) = 75.23, p < 0.0001) but not phonation. There is, however, a significant interaction of phonation and tone (F(3,57) = 9.41, p < 0.0001), which is likely due to the better perception of whispered than phonated T4 as can be seen in the lower panel of Figure 5. This advantage cannot be seen in the original whispered T4 in the upper panel, however. This suggests that it could have been an artifact of the generation of the amplitude-modulated noise.

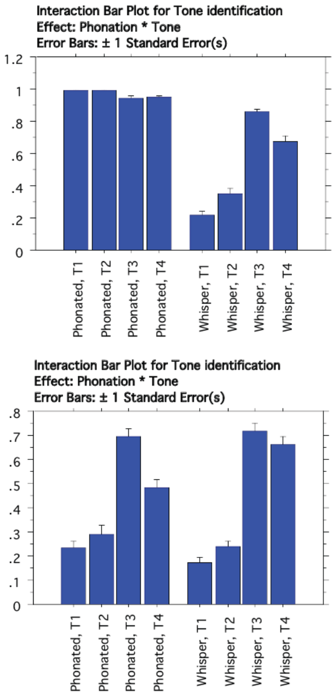

Figure 5: *Tone identification from original speech (upper panel) and amplitude-modulated noise (lower panel) that was phonated (left) or whispered (right).*

Table 2. *Identification rates for phonated and whispered tones in original speech.*

	Heard / Original	T1(%)	T2(%)	T3(%)	T4(%)
Phonated	T1	**98.9**	0.5	0.3	0.3
	T2	0.8	**98.9**	0.0	0.3
	T3	0.0	5.5	**94.2**	0.3
	T4	0.3	0.3	4.7	**94.7**
Whispered	T1	**21.7**	21.1	10.5	46.7
	T2	16.1	**35.8**	26.7	21.4
	T3	0.8	12.2	**85.0**	2.0
	T4	15.6	8.6	8.9	**66.9**

Table 3. *Identification rates for phonated and whispered tones from amplitude-modulated noise.*

	Heard / Original	T1(%)	T2(%)	T3(%)	T4(%)
Phonated	T1	**23.3**	22.5	10.3	43.9
	T2	24.7	**29.2**	18.9	27.2
	T3	5.8	15.0	**69.5**	9.7
	T4	20.3	18.6	12.5	**48.6**
Whispered	T1	**17.2**	22.0	23.3	37.5
	T2	13.0	**23.9**	29.2	33.9
	T3	3.9	18.1	**71.9**	6.1
	T4	12.0	9.7	11.9	**66.4**

More details of the perception tests can be seen in Table 2 and Table 3, which show confusion matrices for tone perception from original and amplitude-modulated tones, respectively. In Table 2 we can see that phonated tones were all perceived accurately, except that T3 and T4 had somewhat lower accuracy. In contrast, T3 is the best perceived in whispered speech. This could be due to its longer duration, relatively low intensity, bi-modal intensity profile, and greater spectral tilt, as shown in Figures 1-4. T1 and T2 were poorly perceived. T1 was heavily confused with both T4 and T2, while T2 is confused with all the other tones.

In Table 3 we can see that T1 and T2 were perceived poorly from amplitude-modulated noise based on both phonated and whispered syllables, with the former slightly better than the latter. T3 and T4 were perceived well above the chance level of 25% for both phonated and whispered syllables, with the latter slightly better than the former. Overall, therefore, once turned into amplitude-modulated noise, phonated tones do not show a clear advantage over whispered tones.

4. Discussion and conclusion

The experimental results presented above have provided answers to the questions raised at the outset of the study. First, the perception of whispered Mandarin tones spoken in isolation is quite good for T3 and T4, and above chance for T2, but is below chance for T1. This is similar to earlier findings [2,15]. Second, acoustic analysis showed a number of possible acoustic cues that varied with tone, including, duration, intensity, temporal profile of intensity and spectral tilt. However, formant frequencies did no show consistent cross-tonal differences. Finally, these cues also existed in the phonated tones, and there was no clear evidence of their enhancement in whispering, as there is lack of interaction between phonation and tone. Therefore, at least from the data of one female speaker, Mandarin does not seem to have developed special strategies to enhance the secondary tonal cues, as claimed by some early studies [2].

This lack of cue enhancement is interesting, given that enhancement has been demonstrated to be possible. [6], for example, has shown that artificially increased spectral tilt helps the perception of whispered boundary tones in Dutch. As seen in Figure 4, however, spectral tilt already varies with Mandarin tones in phonated speech, and the variation of spectral tilt in whispered tones show similar patterns.

In conclusion, although we have found consistent acoustic cues that may have led to the good recognition of some of the Mandarin tones spoken in isolation, we did not find evidence that these cues have been specially developed to enhance whispered tones. Given the limited number of speakers examined for the present paper, this conclusion is rather tentative, and need to be further examined by analyzing more speakers.

5. Acknowledgements

This study was supported in part by the "Chinese Top University Graduate Students Studying Abroad" program from the Chinese Scholarship Council to the first author. Our thanks also go to all the participants at University of Oxford and Tongji University. We also thank Professor John Coleman at University of Oxford for help with recording facility and study design.

6. References

[1] A. S. Abramson, "Tonal experiments with whispered Thai", in A. Valdman [Ed], *Papers on linguistics and phonetics in memory of Pierre Delattre*, pp. 31-44, The Hague: Mouton, 1972.

[2] S. Y. Liu, and A. G. Samuel, "Perception of Mandarin lexical tones when F0 information is neutralized," *Language and Speech*, vol. 47, no.2, pp. 109-138, 2004.

[3] W. Meyer-Eppler, "Realization of prosodic features in whispered speech," *The Journal of the Acoustical Society of America*, vol. 29, pp. 104-106, 1957.

[4] M. Higashikawa, and F. D. Minifie, "Acoustical–perceptual correlates of 'whisper pitch' in synthetically generated vowels," *Journal of Speech, Language, and Hearing Research*, vol. 42, pp. 583–591, 1999.

[5] W. F. L. Heeren, and V. J. Van Heuven, "Perception and production of boundary tones in whispered Dutch," in *Proceedings of Interspeech 2009, September 6-10, Brighton, UK*, 2009, pp. 2411-2414.

[6] W. F. L. Heeren, and V. J. Van Heuven, "The interaction of lexical and phrasal prosody in whispered speech," *The Journal of the Acoustical Society of America*, vol. 136, no.6, pp. 3272-3289, 2014.

[7] D. H. Whalen, and Y. Xu, "Information for Mandarin tones in the amplitude contour and in brief segments," *Phonetica*, vol. 49, pp. 25-47, 1992.

[8] M. Lin, "Putonghua shengdiao de shengxue texing he zhijue zhengzhao [The acoustic characteristics and perceptual cues of tones in Standard Chinese]," *Zhongguo Yuwen* [Chinese Linguistics]，vol. 204, no. 3, pp. 182-193, 1988.

[9] D. Surendran, and G. -A. Levow, "The functional load of tone in Mandarin is as high as that of vowels," in *Proceedings of Speech Prosody 2004*, *March 23-26, Nara, Japan*, 2004, pp. 99-102.

[10] A. D. Patel, Y. Xu, and B.Wang, "The role of F0 variation in the intelligibility of Mandarin sentences," in *Proceedings of Speech Prosody 2010, May 11-14, Chicago, USA*, 2010.

[11] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 12, pp. 117-128, 2000.

[12] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, pp. 341-345, 2001.

[13] Y. Xu, "ProsodyPro — A tool for large-scale systematic prosody analysis," *in Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013), August 30, Aix-en-Provence, France*, 2013, pp. 7-10.

[14] B. Hammarberg, B. Fritzell, J. Gauffin, *et al.*, "Perceptual and acoustic correlates of abnormal voice qualities," *Acta Otolaryngologica*, vol. 90, pp. 441-451, 1980.

[15] M. Gao, *Tones in Whispered Chinese: Articulatory Features and Perceptual Cues*. MA Dissertation, the University of Victoria, 2002.