TONE PRODUCTION IN WHISPERED MANDARIN

Charles Chang and Yao Yao

University of California, Berkeley
cbchang@berkeley.edu, yaoyao@berkeley.edu

ABSTRACT

Acoustic analyses of normal voiced and whispered Mandarin Chinese reveal significant differences in duration and intensity among the four lexical tones, differences that are moreover similar across the two speech genres. In contrast to previous claims, however, these differences among the tones are found to shrink in whisper rather than being exaggerated to facilitate perception. Furthermore, individual variation exists in the production of whispered tones, which are found to shorten or lengthen with respect to normal voiced tones depending on the speaker.

Keywords: Mandarin Chinese, tone, whisper, duration, intensity.

1. INTRODUCTION

Mandarin Chinese has four basic lexical tones (not counting the fifth "neutral" tone), which are known to both native speakers and linguists as Tone 1, Tone 2, Tone 3, and Tone 4. A number of previous studies ([3], [9]) have sketched the fundamental frequency (f_0) contours for the four tones and identified Tone 1 as a high level tone (55), Tone 2 as a rising tone (35), Tone 3 as a low falling rising tone (214), and Tone 4 as a high falling tone (51).

The most important acoustic cue for tone recognition is undoubtedly f_0. The perception of tones when f_0 is absent (e.g. in whispered speech) is thus an interesting problem, though it has not been very widely investigated. Previous studies have not reached a consensus on the accuracy of tone perception in whispered speech (with estimates ranging from 40% to 80%, cf. [5], [6], [7]), but they have pointed to possible cues for tone recognition in whisper. [2], [4], and [8] suggest that temporal envelope and intensity may serve as secondary cues. Moreover, [6] found a correlation between syllable duration and tone perception, which was more significant in human whisper than in machine whisper (processed by removing f_0 information from normal human speech), suggesting that native Mandarin speakers may exaggerate secondary cues such as duration when they know the primary cue to tone is not available (presumably for the benefit of the listener). However, there is no acoustic data in [6] to support this speculation.

This study thus focuses on the acoustics of whispered tones. How do the four Mandarin tones compare to each other acoustically in normal (voiced) speech vs. whisper? When whispering, do speakers exaggerate secondary differences between the tones to make up for the absence of f_0 information? Here we report the results of a production experiment designed to investigate exactly these questions.

2. METHODS

2.1. Stimuli

A list of morphemes was constructed containing all minimal tone quadruplets in which every morpheme would be familiar to a native speaker of Mandarin, excluding those which had: (i) an aspirated plosive or affricate onset, (ii) a non-sibilant fricative onset, (iii) a sonorant onset, or (iv) a syllable coda. These latter syllable types were excluded to make the process of taking acoustic measurements on whispered tokens as straightforward as possible. The resulting list contained a total of 30 tone quadruplets, or 120 morphemes/words.

2.2. Speakers

Two speakers of Beijing Mandarin, LR (female) and DTZ (male), were recorded in this experiment. Both were in their late 20s, and neither reported any history of articulatory or auditory problems.

2.3. Procedure

The speech of both speakers was recorded in a sound-proof booth as mono sound files on a Marantz PMD670 solid state recorder using an AKG C420 head-mounted condenser microphone, which was positioned to the side of the mouth and held at a constant distance of approximately 2 cm from the face. The speakers recorded the same word list in normal speech and whisper, but in a

different random order for each speech genre. For all target words, three tokens were collected in isolation at a sampling rate of 44.1 kHz and a bit rate of 16 bits per sample.

All measurements were taken in Praat 4.5.14 [1] on a Fourier spectrogram with a Gaussian window shape, window length of 5 ms, bandwidth of 200 Hz, dynamic range of 70 dB, and pre-emphasis of 6 dB/octave. Duration was measured from the end of the syllable onset (i.e. the end of the consonant burst or strident interval) to the end of visible formant structure in the vowel, and average intensity was measured over this vowel interval.

Example spectrograms and intensity contours for the [ta] tone quadruplet (all from Speaker DTZ, token 2) are presented below in Figures 1-4. Note the similarities in formant structure between normal speech and whisper, as well as the differences among the intensity contours of the four tones. Tones 1 and 2 have nearly identical, relatively flat contours in both normal speech and whisper; Tone 3 dips in intensity in the middle; and Tone 4 begins to drop off in intensity relatively early.

Figure 1: Spectrograms and intensity contours of 答 [ta^{55}] 'to answer' in normal speech (L) and whisper (R).

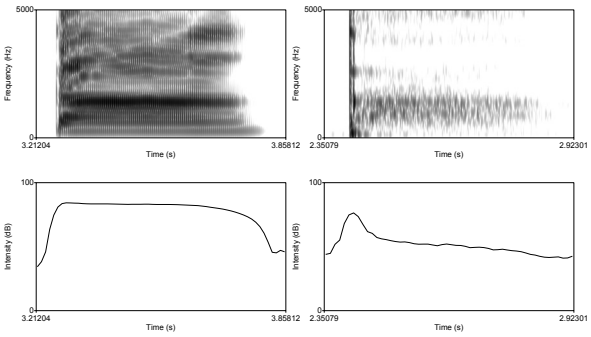


Figure 2: Spectrograms and intensity contours of 达 [ta^{35}] 'to reach' in normal speech (L) and whisper (R).

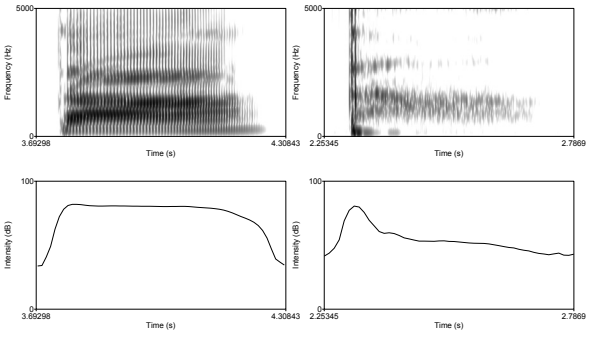


Figure 3: Spectrograms and intensity contours of 打 [ta^{214}] 'to hit' in normal speech (L) and whisper (R).

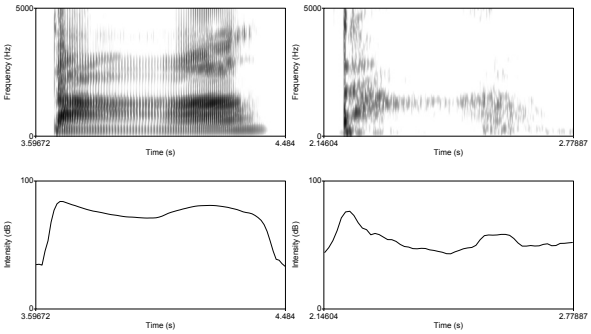
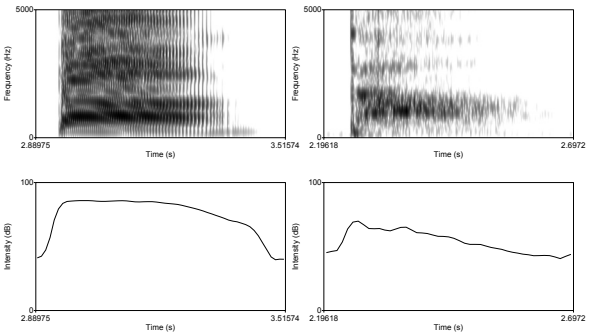


Figure 4: Spectrograms and intensity contours of 大 [ta^{51}] 'big' in normal speech (L) and whisper (R).



3. RESULTS

3.1. Duration

The relative durations of Tones 1-4 found here conform to descriptions in previous research (e.g. [9]). Tone 3 is the longest and Tone 4 the shortest, with Tones 1 and 2 being intermediate in duration.

Table 1: Results of pair-wise comparisons between tone durations in normal speech.

Comparison	*t*	df	*p*-value
Tone 1 vs. Tone 2	4.488	59	< .0005
Tone 1 vs. Tone 3	-14.289	59	< .0005
Tone 1 vs. Tone 4	15.478	59	< .0005
Tone 2 vs. Tone 3	-13.137	59	< .0005
Tone 2 vs. Tone 4	9.385	59	< .0005
Tone 3 vs. Tone 4	25.364	59	< .0005

Table 2: Results of pair-wise comparisons between tone durations in whisper.

Comparison	*t*	df	*p*-value
Tone 1 vs. Tone 2	1.500	59	n.s.
Tone 1 vs. Tone 3	-10.290	59	< .0005
Tone 1 vs. Tone 4	7.568	59	< .0005
Tone 2 vs. Tone 3	-12.424	59	< .0005
Tone 2 vs. Tone 4	7.357	59	< .0005
Tone 3 vs. Tone 4	15.240	59	< .0005

All pair-wise duration differences between the tones in normal speech are highly significant. With the exception of Tone 1 vs. Tone 2, the duration differences in whisper are also highly significant. The results of paired-samples *t*-tests are given in Tables 1 and 2 above.

A repeated-measures analysis of variance (ANOVA) shows a main effect of tone (Speaker LR: $F(3, 87) = 177.467$, $p < .0005$; Speaker DTZ: $F(3, 87) = 302.799$, $p < .0005$); a main effect of genre (Speaker LR: $F(1, 29) = 37.588$, $p < .0005$; Speaker DTZ: $F(1, 29) = 123.033$, $p < .0005$); and a [tone x genre] interaction for both speakers (Speaker LR: $F(3, 87) = 4.317$, $p = .007$; Speaker DTZ: $F(3, 87) = 22.931$, $p < .0005$).

Average tone durations across words for each speaker are presented in Figures 5 and 6 below. Note that relative durations of the tones with respect to each other are the same across speakers (Tone 3 being the longest, followed by Tones 1 and 2 in either order, then by Tone 4) and remain the same across the different genres.

Figure 5: Average tone duration across words for Speaker LR.

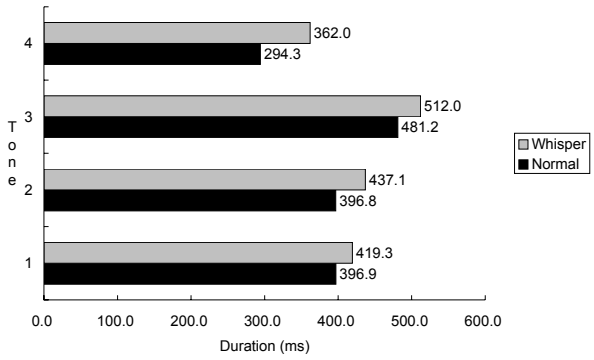
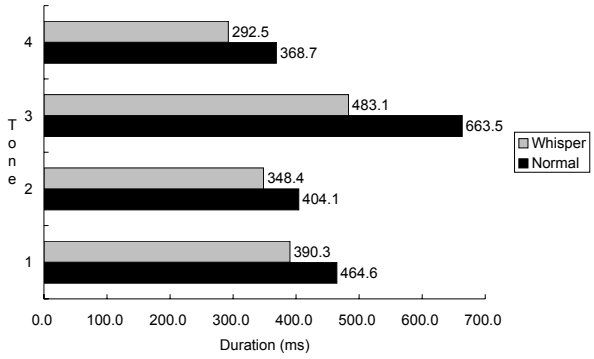


Figure 6: Average tone duration across words for Speaker DTZ.



Two features of Figures 5 and 6 should be noted in particular. First, neither speaker appears to have exaggerated the tonal duration differences in whisper as compared to normal speech. On the contrary, the four tones are more similar to each other in whisper, as seen in the lower standard deviations of average durations in whisper (Speaker LR: 61.9 ms vs. 76.5 ms; Speaker DTZ: 80.4 ms vs. 131.6 ms). Second, there are significant individual differences between speakers with respect to tone durations across the different genres. Speaker LR consistently lengthens tones in whisper as compared to normal speech, while Speaker DTZ consistently shortens them.

Thus, these results do not support the idea that speakers promote the duration cue when whispering for the benefit of listeners. Variation in duration across different speech genres appears to be related to individual speech style and might not have a consistent, speaker-independent pattern.

3.2. Intensity

In normal speech, Tone 4 has the highest average intensity and Tone 3 the lowest, with Tone 1 and Tone 2 being intermediate in intensity. All pair-wise intensity differences in normal speech are highly significant. In whisper, however, the differences between the tones shrink, and the only significant difference that remains is that between Tone 1 and Tone 3. The results of paired-samples *t*-tests are given in Tables 3 and 4 below.

Table 3: Results of pair-wise comparisons between average tone intensities in normal speech.

Comparison	*t*	df	*p*-value
Tone 1 vs. Tone 2	2.486	59	.016
Tone 1 vs. Tone 3	11.450	59	< .0005
Tone 1 vs. Tone 4	5.512	59	< .0005
Tone 2 vs. Tone 3	12.197	59	< .0005
Tone 2 vs. Tone 4	3.943	59	< .0005
Tone 3 vs. Tone 4	-8.863	59	< .0005

Table 4: Results of pair-wise comparisons between average tone intensities in whisper.

Comparison	*t*	df	*p*-value
Tone 1 vs. Tone 2	1.728	59	n.s.
Tone 1 vs. Tone 3	3.869	59	< .0005
Tone 1 vs. Tone 4	-0.370	59	n.s.
Tone 2 vs. Tone 3	1.877	59	n.s.
Tone 2 vs. Tone 4	-1.012	59	n.s.
Tone 3 vs. Tone 4	-1.932	59	n.s.

A repeated-measures analysis of variance (ANOVA) again shows main effects of tone (Speaker LR: $F(3, 87) = 38.551$, $p < .0005$; Speaker DTZ: $F(3, 87) = 3.535$, $p = .018$) and genre (Speaker LR: $F(1, 29) = 420.684$, $p < .0005$; Speaker DTZ: $F(1, 29) = 928.238$, $p < .0005$) for

both speakers, and a [tone x genre] interaction for Speaker LR ($F(3, 87) = 24.598$, $p < .0005$).

Average tone intensities across words for each speaker are presented in Figures 7 and 8 below. As with relative durations, relative intensities are the same across speakers. In normal speech, Tone 1 has the highest average intensity (followed by Tone 2, Tone 4, and then Tone 3), whereas in whisper, Tone 4 has the highest average intensity.

Figure 7: Average tone intensity across words for Speaker LR.

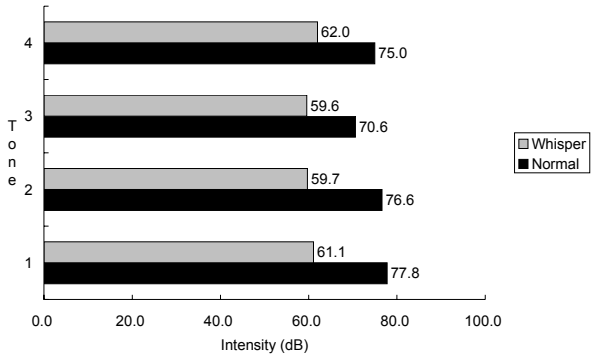
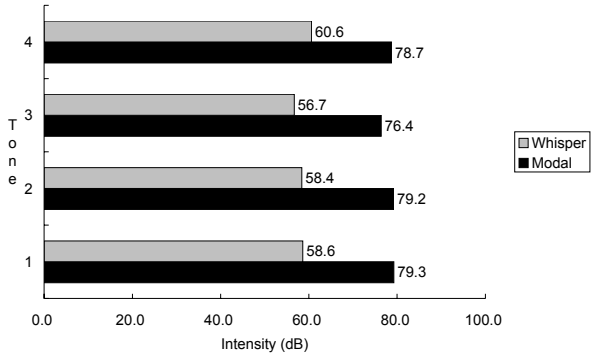


Figure 8: Average tone intensity across words for Speaker DTZ.



As with duration, neither speaker appears to have exaggerated the tonal intensity differences in whisper as compared to normal speech. If anything, the four tones are more similar in whisper than in normal speech, especially for Speaker LR. Not unexpectedly, for both speakers average intensity of all four tones decreases significantly in whisper as compared to normal speech.

4. CONCLUSION

To summarize, we recorded two native Mandarin speakers' productions of isolated syllables in both normal speech and whisper and examined two variables – duration and average intensity – in their productions. Contrary to what [6] would predict, for both speakers the differences in duration and intensity among the four tones do not become larger in whisper. In fact, duration and intensity are more similar across tones in whisper than in normal speech. These findings indicate that speakers are not exaggerating cross-tonal duration or intensity differences in whisper and, thus, that listeners do not have any more duration or intensity information available to them for recognizing tones in whisper as they do in normal speech. The perception results of [6], then, are presumably due to some degree of unnaturalness of the "machine whispered" stimuli rather than exaggerated secondary cues in the "human whispered" stimuli.

As mentioned in the introduction, this study is part of ongoing research on tone in whisper. Even with two speakers, variation was found in duration differences across normal speech and whisper. Current work is examining whether this variation is idiosyncratic or correlated with other factors like gender and dialectal background.

Given that differences in duration and intensity among tones are diminished in whisper, but still quite significant, it is an interesting question whether listeners make use of these muted secondary cues in the absence of f_0. Thus, future research will examine the effect of duration, average intensity, and intensity contour on tonal identification in whisper in perception experiments with both native Mandarin speakers as well as second-language learners of Mandarin.

5. REFERENCES

[1] Boersma, P., Weenink, D. 2007. Praat: doing phonetics by computer. http://www.praat.org.

[2] Fu, Q.-J., Zeng, F.-G., Shannon, R.V., Soli, S.D. 1998. Importance of tonal envelope cues in Chinese speech recognition. *J. Acoust. Soc. Am.* 104, 505-510.

[3] Gandour, J. 1978. The perception of tone. In: V.A. Fromkin (ed), *Tone: A Linguistic Survey*.

[4] Kong, Y.-Y., Zeng, F.-G. 2006. Temporal and spectral cues in Mandarin tone recognition. *J. Acoust. Soc. Am.* 120.5, 2830-2840.

[5] Liang, Z.-A. 1963. Hanyu putonghua zhong shengdiao de tingjiao bianren yiju (The auditory basis of tone recognition in Standard Chinese). *Acta Phys. Sin.* 26, 85-91.

[6] Liu, S., Samuel, A. 2004. Perception of Mandarin lexical tones when F0 information is neutralized. *Language and Speech* 47.2, 109-138.

[7] Miller, J.D. 1961. Word tone recognition in Vietnamese whispered speech. *Word* 17, 11-15.

[8] Nicholson, H., Teig, A.H. 2003. How to tell beans from farmers: cues to the perception of pitch accent in whispered Norwegian. *Nordlyd* 31.2, 315-325.

[9] Xu, Y. 1997. Contextual tonal variations in Mandarin. *J. Phonetics* 25, 61-83.