Whispered Mandarin has no production-enhanced cues for tone and intonation

Abstract

It is often assumed, explicitly or implicitly, that speakers generate special cues in whispered tone and intonation to make up for the absence of fundamental frequency. The present study examined this assumption with one production and three perception experiments. The production experiment compared duration, intensity, formants and spectral tilt of phonated and whispered Mandarin monosyllabic utterances with four lexical tones spoken as either statements or questions. For tones, no acoustic properties were found to occur only in whispered but not in phonated utterances. For intonation, some spectral tilt measurements differed between the two phonation types. The two tone perception experiments used phonated and whispered utterances as well as amplitude-modulated noise based on those utterances as stimuli. Results show that once turned into amplitude-modulated noise, phonated and whispered tones had similar identification patterns, indicating that the non-F_0 tonal cues in whispers were already in phonated speech. The intonation perception experiment used original utterances as stimuli and showed a substantial drop in overall identification rate and an overwhelming bias toward statement. Thus the spectral tilt differences found in the acoustic analysis were not helpful for intonation perception. Possible reasons for the lack of effective enhancement in whispered speech were discussed.

Keywords: special cues; whispered Chinese tones; whispered Chinese intonations

1. Introduction

In whispered speech, the periodic voice source during vocalic sounds is replaced by noise sources (Laver, 1994; Zemlin, 1988). As a result, there is no fundamental frequency

in the utterance. This creates a problem for tone and intonation, for which the major cues are carried by pitch (Yip, 2002). Yet there is evidence that listeners can still perceive some pitch (Thomas, 1969; Heeren, 2015), tones (Wise and Chong, 1957; Abramson, 1972; Gao, 2002) in tone languages, boundary tones (Heeren and Van Heuven, 2014) in a non-tone language, and intonation (Żygis *et al*., 2017) from whispers. There must therefore be non-F_0 cues in the speech signal that can compensate for the absence of fundamental frequency in whispered utterances. What has not been clearly demonstrated, however, is whether the non-F_0 cues are produced only for whispers, or they are already present in phonated speech but become more useful in the absence of F_0.[1] The first mechanism can be referred to as the *production enhancement* account, while the second the *perceptual compensation* account. Based on the production enhancement account, speakers have developed special articulatory strategies to make up for the absence of F_0 so as to aid listeners to perceive tone and intonation in whispers. As suggested by a number of studies, implicitly or explicitly, certain non-melodic properties are modified by speakers to compensate for the loss of melodic cues. Meyer-Eppler (1957) states that there are two substitutes for pitch in whispered vowels: spectral noise ([e], [i], and [o]) and formant position change ([a] and [u]). Gao (2002) concludes that there are two special maneuvers for Tone 3 and Tone 4 in Chinese: a) males lengthen the vocalic duration and b) females exaggerate the amplitude contours. Similarly, Liu and Samuel (2004) conclude that "Mandarin speakers promote the utility of secondary cues when they know that the primary cue will be unavailable" (p. 109), and that "native Mandarin speakers, when required to produce monosyllables without their primary cue to tone identity, increased the salience of secondary cues" (p. 132). Recently, Heeren (2015) shows that formants, center of gravity, spectral balance and intensity are different with low, mid and high pitch targets in whispered and normally phonated utterances, suggesting that speakers can develop a compensatory strategy in whispers. She further demonstrates with acoustic manipulations that such a compensatory strategy would be beneficial for perceiving whispered intonation. Żygis *et al*. (2017:53) further

[1] Note that a cue is not the same as an acoustic dimension such as intensity, duration, spectral tilt or formants. Rather, a cue is a special value of an acoustic dimension used to mark a particular phonological or phonetic contrast. For example, a High tone is cued by a *high* F_0 rather than merely by the acoustic dimension of F_0.

suggest that “speakers produce intended intonation patterns by varying the type and magnitude of cues depending on speech mode”, so as to make some cues more pronounced in whispered than in phonated speech (p. 69).

The perceptual compensation account, in contrast, assumes that the non-pitch tonal cues are already present in phonated speech, but they become more useful only in the absence of F_0 in the case of whispers. Abramson (1972), for example, shows that concomitant tonal features are present in phonated speech, and that when voicing is removed in whispers, they become more audible. He further points out that the amount of ambiguity due to the loss of F_0 in whispers is a function of utterance length: the longer the utterance, the less the ambiguity. Chang and Yao (2007) find that both normal voiced and whispered Mandarin Chinese show similar differences in duration and intensity among the four lexical tones; but those differences are actually reduced rather than exaggerated in whispers. It has also been shown that in phonated speech, the role of pitch variation is so dominant that the effect of other phonetic cues on the identification of tones can be hardly demonstrated (Abramson, 1972; Lin, 1988). It is only when F_0 is absent, e.g., in signal-correlated noise, that the role of amplitude profiles and duration becomes discernible (Whalen and Xu, 1992). Thus the increased relevance of the non-F_0 cues is viewed only as a perceptual phenomenon in the perceptual compensation account.

The key difference between the two accounts is therefore whether the non-F_0 cues in whispers come from special articulatory manoeuvres developed by speakers to compensate for the absence of fundamental frequency. These manoeuvres are either absent in phonated speech, or they are exaggerated in whispers. Also, these manoeuvres are not made for the production of whispers *per se*, but aimed directly at facilitating pitch perception in the absence of F_0.

The present study is a further exploration of whispered speech in Mandarin, with the aim to determine whether speakers have developed special acoustic cues to aid the perception of tone and intonation in the language. For this purpose, one production and three perception experiments were conducted.

2. Materials and Method

Table 1 shows the production targets as well as the perception stimuli for the current study.[1] There are five sets of syllables composed of only vowels (/a/, /ɤ/, and /u/) or glide onsets (/i/-yi and /y/-yü). The use of vowel-only syllables was to ensure the inclusion of full tonal contours, based on the assumption that a tone is carried by the entire syllable rather than just the rhyme (Xu and Xu, 2003; Xu, 2004). These syllables also would lead to the least undesirable artefact in generating amplitude-modulated noise in one of the perception experiments. Words with the same CV syllable are further distinguished by four lexical tones (T1: High level, T2: Rising, T3: Low, and T4: Falling). In terms of intonation, only statement and its corresponding echo question are considered in this project.

Table 1. *A list of syllables for production and perception experiments.*

Tone \ Vowel		a	ɤ	i	u	y
T1	Character	啊	婀	衣	乌	迂
	Pinyin	*ā*	*ē*	*yī*	*wū*	*yū*
	Glossary	*oh*	*graceful*	*clothes*	*black*	*winding*
T2	Character	啊	鹅	姨	无	鱼
	Pinyin	*á*	*é*	*yí*	*wú*	*yú*
	Glossary	*eh*	*goose*	*aunt*	*nothing*	*fish*
T3	Character	啊	恶	椅	五	雨
	Pinyin	*ǎ*	*ě*	*yǐ*	*wǔ*	*yǔ*
	Glossary	*what*	*nausea*	*chair*	*five*	*rain*
T4	Character	啊	饿	意	物	玉
	Pinyin	*à*	*è*	*yì*	*wù*	*yù*
	Glossary	*ah*	*hungry*	*meaning*	*thing*	*jade*

[1] The list is a sub-list from a larger project (Jiao *et al*, 2015; Jiao and Xu, 2016).

2.1. Production Experiment

2.1.1. Subjects

Twelve native Mandarin speaking students (average age: 20.3; six females) participated in the recording session. They were divided into six pairs with one male and one female in each pair. One pair did the recording at University of Oxford and the other five at Tongji University. None of the participants reported any impairment in speech, hearing or vision. They were given a small compensation for their time and effort.

2.1.2. Stimuli

All the characters in Table 1 were used as production stimuli. Each character was spoken both as a statement and as a question under both phonated and whispered conditions. Overall, there were a total of 80 stimuli (5 syllables * 4 tones * 2 intonations * 2 phonation types) by each speaker.

2.1.3. Recording Procedure

Each pair of speakers sat side by side in the recording booth in the lab, where target characters with the corresponding pinyin were presented on the computer screen. They were asked to perform a dialogue, with one saying the character as a question and the other saying the same character as an answer. And then for the next character, the pair reversed their roles and the second speaker became the one who asked the question. Illustrations are given in Table 2. This paradigm ensured that the speakers were made constantly aware of the communication intent of both tone and intonation. The phonated and whispered dialogues were done in separate blocks, and the phonated block always preceded the whispered block for each pair of speakers. A microphone was put on a stand between the two speakers, with a distance of around 15 cm from each. The input volume was set to be the same for phonated and whispered conditions, which was neither too loud for the phonated nor too soft for the whispered. The experimenter monitored and controlled the progression of the recording outside the booth. At the Oxford site, the recording was done with an Audio-Technica AT4031 microphone and the sounds were recorded onto a compact disk by a CD recorder (HHB CDR-850) at 44.1 kHz and 16 bits resolution, and transferred into a PC using a Sound Blaster analogue to digital conversion. At the Tongji site, a

Neumann U87 microphone was used. The audio was recorded by Pro Tools 8.1 and saved in .wav form at 44.1 kHz and 24 bits resolution.

Table 2. *Examples of how a pair of speakers spoke each character in two intonations and how they rotated their roles for the next character (The first two columns).*

Speaker 1:	yǔ?	*"rain?"*	*(question)*
Speaker 2:	yǔ.	*"rain."*	*(statement)*
Speaker 2:	è?	*"hungry?"*	*(question)*
Speaker 1:	è.	*"hungry."*	*(statement)*

2.2. Perception Experiments

2.2.1. Subjects

Twenty-two native mandarin students (average age: 20.2; twelve females) were recruited for the perception experiments. They had no self-reported speech or hearing disorders. They were also given a small compensation for their time and effort.

2.2.2. Stimuli

Phonated and whispered characters from three vowel columns in Table 1 (/ɤ/, /i/ and /y/) in both intonations by the female speaker from the Oxford pair were used as stimuli for the first two experiments (tone and intonation identification from the original speech). For the third experiment (tone identification from amplitude-modulated noise), these natural tokens were used as the base to generate amplitude-modulated noise. A Praat (Boersma, 2001) script was written to first extract the Amplitude tier of base stimuli, with the duration of the base also retained. The amplitude profile was then imposed onto a pink noise of the same duration (generated in the script by filtering white noise with a -6 dB/octave de-emphasis starting from 50 Hz). The sound generated this way preserved the original duration and amplitude profile, with no spectral information left. 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 in overall intensity.

Overall, there were 96 trials for each listener (3 syllables * 4 tones * 2 intonations * 2 phonation types * 2 repetitions).

2.2.3. Procedure

Participants took part in the perception experiments in a quiet room in Tongji University. The tests were run with Praat through ExperimentMFC scripts. Subjects wore Sennheiser PC 166 headphones and sat in front of a Dell computer (OPTIPLEX 390) in a confortable position. In each trial, the subject heard an utterance (or noise), and saw four Chinese characters of the corresponding syllables with four tones on the screen. They then selected the character closest to what they had heard. Each stimulus was played only once.

It took around 30-40 minutes for each subject to finish. All of them were tested first with the natural speech stimuli (tone and intonation identifications), and then with the amplitude-modulated noise (tone identification). For the natural speech tasks, around half (10) heard the whispered block first, and the other half heard the phonated block first. Within each block, the stimuli were randomized. And the randomization was different for each participant.

3. Results

3.1. Acoustic Analysis and Results

We extracted formants for production analysis using FormantPro (Xu, 2007-2015) and ProsodyPro (Xu, 2013), which are Praat scripts for large-scale spectral and prosodic analysis. Table 3 shows the major measurements generated by the two scripts that were subjected to further analysis. These measurements were analyzed by four-way Repeated Measures ANOVAs (intonation, phonation, tone, and vowel).

Table 3. *Acoustic Measurements.*

Duration (ms)	Duration of target syllable
Intensity (dB)	Mean intensity of target syllable
Spectral Center of Gravity (COG) (Hz)	Center of spectral gravity

Hammarberg Index (dB)	Difference between the energy in the 0-2kHZ and 2-5kHz bands (Hammarberg *et al*, 1980)
Energy below 500 Hz (dB)	Energy of voiced segments below 500 Hz
Energy below 1000 Hz (dB)	Energy of voiced segments below 1000 Hz
F1, F2, F3 (Hz)	Frequencies of first three formants at syllable center

3.1.1. Effects of Phonation Type

Table 4 lists significant results from the four-way Repeated Measures ANOVAs. First, there is a main effect of phonation for all acoustic measurements. Compared to phonated utterances, whispers had longer duration (508.743 vs. 384.664 ms), weaker intensity (41.056 vs. 66.005 dB), higher center of gravity (1228.753 vs. 547.657 Hz), less difference in energy between 0-2kHz and 2-5kHz (Hammarberg Index: 11.573 vs. 21.24 dB), less energy below 500 Hz (0.349 vs. 0.679 dB) and below 1000 Hz (0.543 vs. 0.858 dB), higher F1 (779.443 vs 511.038 Hz), F2 (1821.279 vs. 1524.67 Hz) and F3 (3105.766 vs. 2759.81 Hz).

Table 4. *Significant effects by four-way Repeated Measures ANOVAs.*

Measurements	Variables	DF	F-Value	P-Value
Duration	intonation	1,11	17.121	0.0017
	phonation	1,11	96.664	<0.0001
	tone	3,33	46.651	<0.0001
	vowel	4,44	16.233	<0.0001
	intonation * tone	3,33	30.155	<0.0001
	phonation * tone	3,33	3.53	0.0253
Intensity	intonation	1,11	16.92	0.0017
	phonation	1,11	586.61	<0.0001
	tone	3,33	18.857	<0.0001
	vowel	4,44	69.726	<0.0001

	intonation * tone	3,33	3.047	0.0423
	intonation * vowel	4,44	6.131	0.0005
	phonation * tone	3,33	14.389	<0.0001
	phonation * vowel	4,44	18.613	<0.0001
	tone * vowel	12,132	3.989	<0.0001
COG	intonation	1,11	10.107	0.0088
	phonation	1,11	57.474	<0.0001
	vowel	4,44	26.014	<0.0001
	intonation * phonation	1,11	8.231	0.0153
	phonation * vowel	4,44	23.878	<0.0001
Hammarberg Index	phonation	1,11	60.699	<0.0001
	vowel	4,44	172.773	<0.0001
	intonation * phonation	1,11	9.557	0.0103
	phonation * tone	3,33	4.144	0.0134
	phonation * vowel	4,44	13.212	<0.0001
	tone * vowel	12,132	2.351	0.0091
Energy below 500 Hz	intonation	1,11	12.942	0.0042
	phonation	1,11	87.338	<0.0001
	tone	3,33	3.68	0.0217
	vowel	4,44	208.306	<0.0001
	intonation * vowel	4,44	2.723	0.0414
	phonation * tone	3,33	3.759	0.0199
	phonation * vowel	4,44	27.166	<0.0001
	intonation * phonation * vowel	4,44	3.426	0.016
	phonation * tone * vowel	12,132	3.472	0.0002
	intonation * phonation * tone * vowel	12,132	3.079	0.0007
Energy below 1000 Hz	phonation	1,11	72.882	<0.0001

	tone	3,33	3.103	0.0398
	vowel	4,44	119.511	<0.0001
	intonation * phonation	1,11	5.877	0.0337
	phonation * vowel	4,44	20.836	<0.0001
	phonation * tone * vowel	12,132	2.555	0.0045
F1	phonation	1,11	77.139	<0.0001
	tone	3,33	8.254	0.0003
	vowel	4,44	170.049	<0.0001
	phonation * vowel	4,44	6.064	0.0006
F2	phonation	1,11	106.76	<0.0001
	vowel	4,44	443.37	<0.0001
	phonation * tone	3,33	3.574	0.0242
	phonation * vowel	4,44	21.6	<0.0001
	tone * vowel	12,132	4.183	<0.0001
	intonation * tone * vowel	12,132	2.276	0.0117
	phonation * tone * vowel	12,132	2.465	0.0062
F3	phonation	1,11	230.809	<0.0001
	vowel	4,44	39.157	<0.0001
	intonation * vowel	4,44	7.695	<0.0001
	phonation * vowel	4,44	8.993	<0.0001
	tone * vowel	12,132	3.719	<0.0001

3.1.2. Effect of tone and its interaction with phonation

Figures 1-2 and 4-5 present bar graphs of significant effects (i.e., those shown in Table 4) of tone and its interaction with phonation. The interactions, in particular, allow us to assess if there is any “enhancement” in whispered tones. Figure 1a shows mean durations of the four tones and their standard errors, averaged across both phonated and whispered conditions. The duration of T3 is the longest, and Student-Newman-Keuls tests found T3 to be significantly longer than all the other tones, but there were no significant duration

differences among the other tones. Figure 1b shows that, from phonated to whispered, the durations of all tones were lengthened, but the overall patterns remained, i.e., T3 is the longest, while the other tones have similar durations. Two separate 3-way repeated measures ANOVAs showed that there were significant effects of tone on duration for both phonated ($F(3,33) = 81.50$, $p < 0.0001$) and whispered ($F(3,33) = 17.97$, $p < 0.0001$) utterances. And Student-Newman-Keuls tests showed significant differences between T3 and all the other three tones in both phonated and whispered utterances, but not between any other two tones. Thus there is no evident enhancement for whispers in terms of duration.

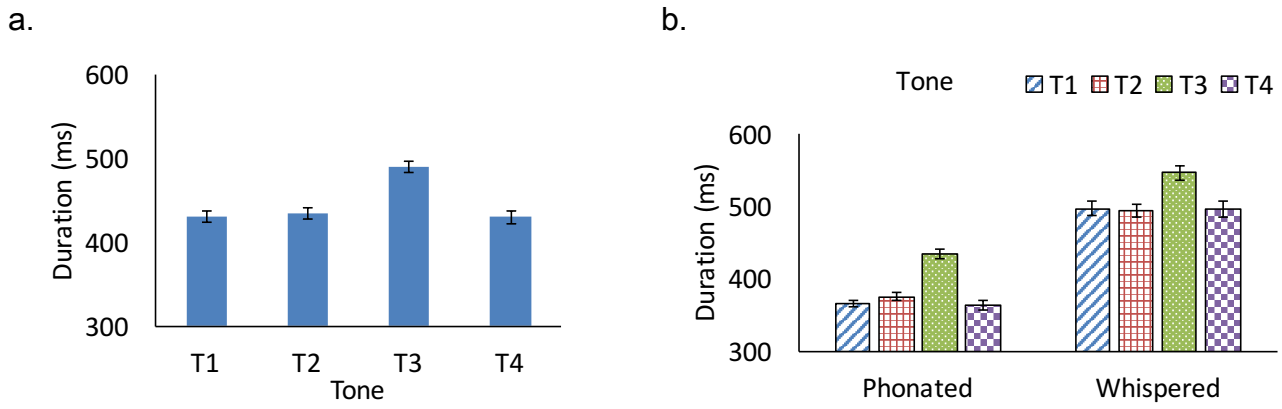


Figure 1: *Mean duration as a function of tone (a) and phonation types and tones (b), with standard errors.*

Figure 2a displays mean intensity of the four tones and their standard errors. T4 shows the greatest intensity and T3 the smallest. Student-Newman-Keuls tests show that the intensity of T3 is significantly lower than all the other three tones, and that the intensity of T4 is greater than that of T2. Figure 2b shows that, first, intensity is overwhelmingly weaker in whispers than in phonated utterances, and secondly, intensity pattern of tones in phonated conditions is in line with that in Figure 2a. Two separate 3-way repeated measures ANOVAs showed that there was a significant effect of tone on intensity only for phonated utterances ($F(3,33) = 38.53$, $p < 0.0001$). And Student-Newman-Keuls tests showed significant differences in all in tone pairs except T1-T4 in phonated utterances, but no difference in any tone pairs in whispered utterances. Thus the tonal differences in intensity are much reduced in whispered utterances.

Figure 3 displays averaged temporal profiles of intensity of the four tones in phonated as well as whispered utterances. As can be seen, for each tone, the profiles appear similar between the two phonation types. Across the tones, T3 stands out with an extra long duration and bimodal profile, and T4 with a short duration and slightly greater drop in the later half of the syllable. These two properties have been found to be responsible for the better perception of these two tones from signal-correlated noise in Author and Author (1992), where the amplitude profiles and duration patterns are both from phonated utterances only. Figure 3 here shows further that no enhancement of intensity and duration cues for the tones are generated in whispered utterances.

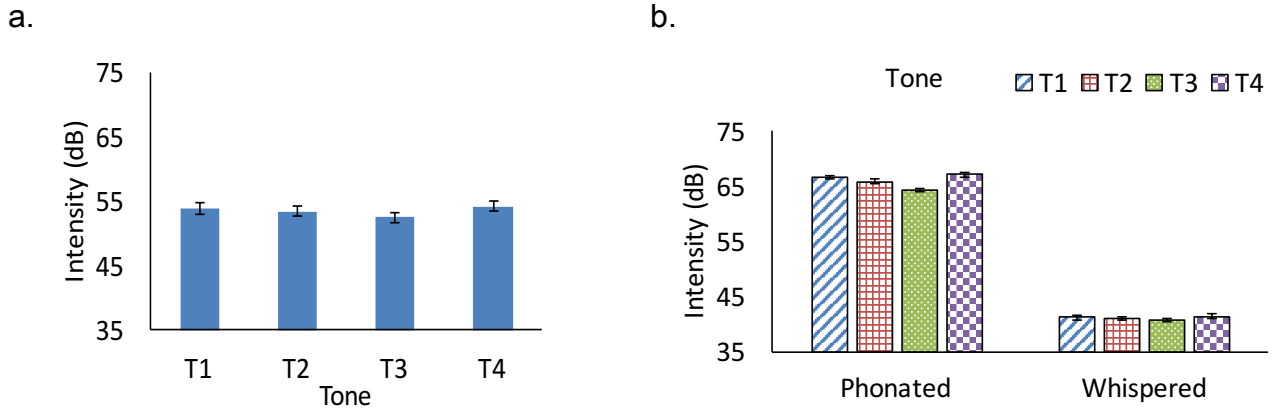
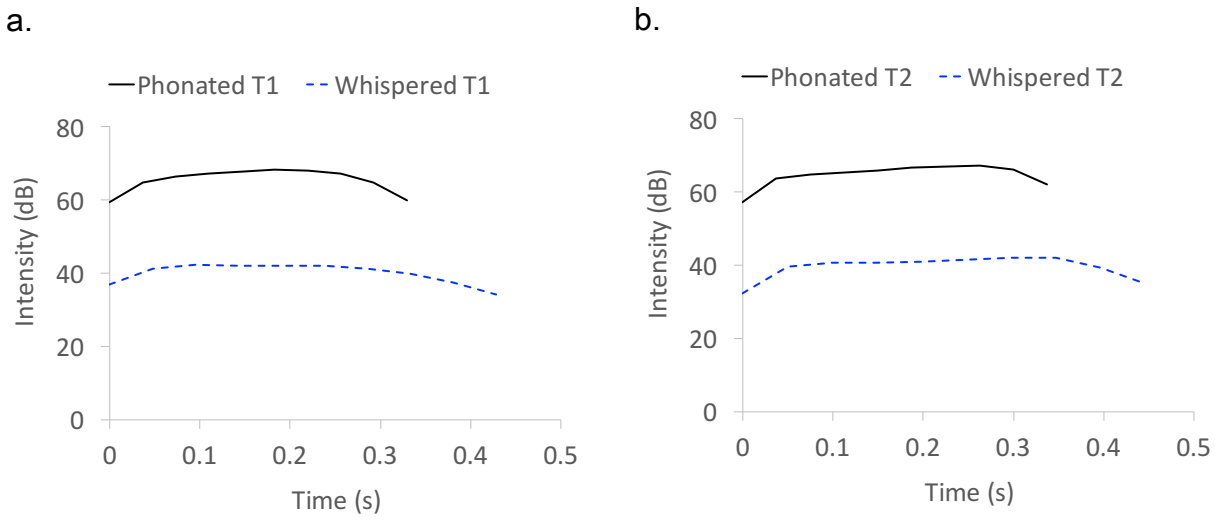


Figure 2: *Mean intensity as a function of tone (a) and phonation type and tone (b), with standard errors.*



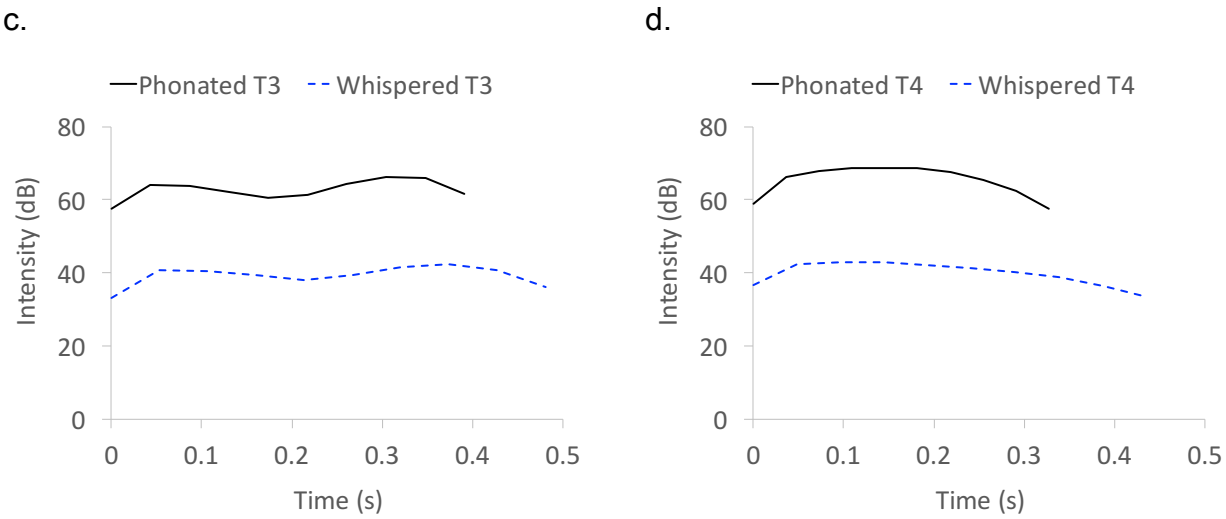


Figure 3: *Intensity profile as a function of time for T1 (a), T2 (b), T3 (c) and T4 (d), respectively.*

Figure 4a shows mean values of energy below 500 Hz, an indicator of spectral tilt, for which there was a marginal overall effect of tone as shown in Table 4. A Student-Newman-Keuls test only shows a significantly greater value in T2 than T3, as can be also seen in the bar graph. In Figure 4b, one can see that there is a lessening of energy in whispered utterances. Two separate 3-way repeated measures ANOVAs showed that there was a significant effect of tone on energy below 500 Hz only for phonated ($F(3,33) = 4.71$, $p = 0.0076$), but not for whispered utterances. And Student-Newman-Keuls tests showed significant differences between T2 and all the other three tones in phonated utterances but not in whispered utterances. There was only a significant difference between T1 and T3 in whispers. Thus again, there is actually a reduction of differences in this measurement across the four tones in the whispered condition.

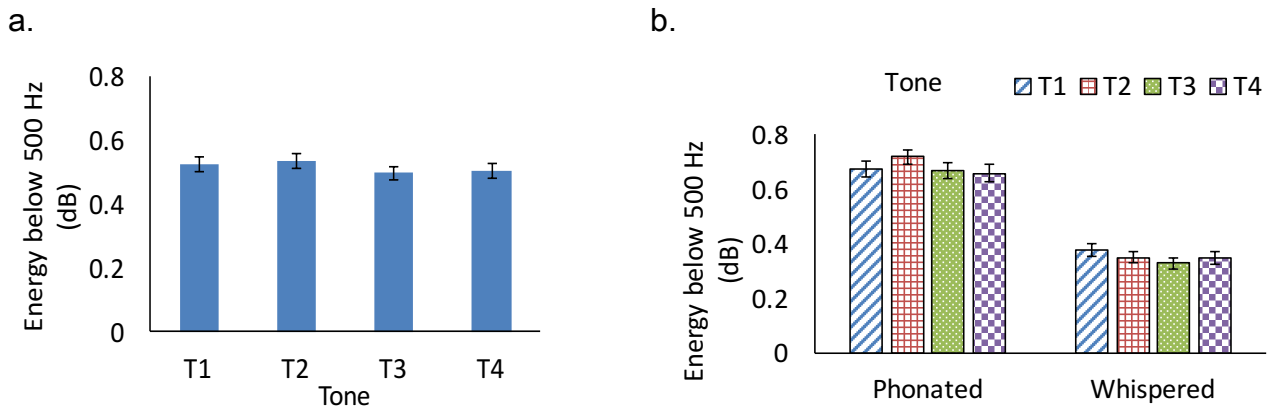


Figure 4: *Mean values of energy below 500 Hz, as a function of tone (a) and phonation type and tone (b), with standard errors.*

As shown in Table 4, there were significant interactions of phonation and tone on Hammarberg Index and F2 although there were no main effects on these two measurements. In Figure 5a, one can see that there is not only an overall reduction of Hammarberg Index for four tones from phonated to whispered utterances, but also a change in the difference across the tones. Two separate 3-way repeated measures ANOVAs showed that there was a significant effect of tone on Hammarberg Index only for phonated ($F(3,33) = 3.84$, $p = 0.0183$), but not for whispered utterances. Student-Newman-Keuls tests showed significant differences only between T2 and T4 in phonated utterances but no differences between any two tones in whispered utterances.

In Figure 5b, there are changes in both overall and cross-tone distribution of F2 values from phonated to whispered utterances. Two separate 3-way repeated measures ANOVAs showed that there was a significant effect of tone on F2 only for whispered ($F(3,33) = 3.81$, $p = 0.0189$), but not for phonated utterances. A Student-Newman-Keuls test showed that T3 had significant higher F2 than T2 and T4 only in whispered utterances. It is curious, however, why F2 would be raised in T3 relative to other tones, given that pitch lowering often involves lowering of the larynx (Honda *et al.*, 1999; Moisik *et al.*, 2014), which should have led to decreased rather than increased formants (due to lengthened vocal tract). To examine what may be the cause, we plotted continuous F2 trajectories by all speakers in phonated and whispered utterances, as shown in Figure 6. As can be seen in Figure 6b, F2 of T3 deviates from the other three tones in the middle section of the whispered syllable, which is absent in the phonated syllables in Figure 6a. On an even closer look, 8 out of the 12 speakers

showed the upward bulge in F2 in the middle section of the syllable, but other 4 speakers did not. Interestingly, 2 of the speakers also showed similar F2 deviations in T2, which also has a dip in its underlying pitch trajectory. Thus the significant raising of F2 in T3 is not very likely due to an enhancement maneuver, but to formant tracking errors related to low-pitch articulation already occurring in phonated speech. The exact cause of the errors, however, needs to be investigated in future research.

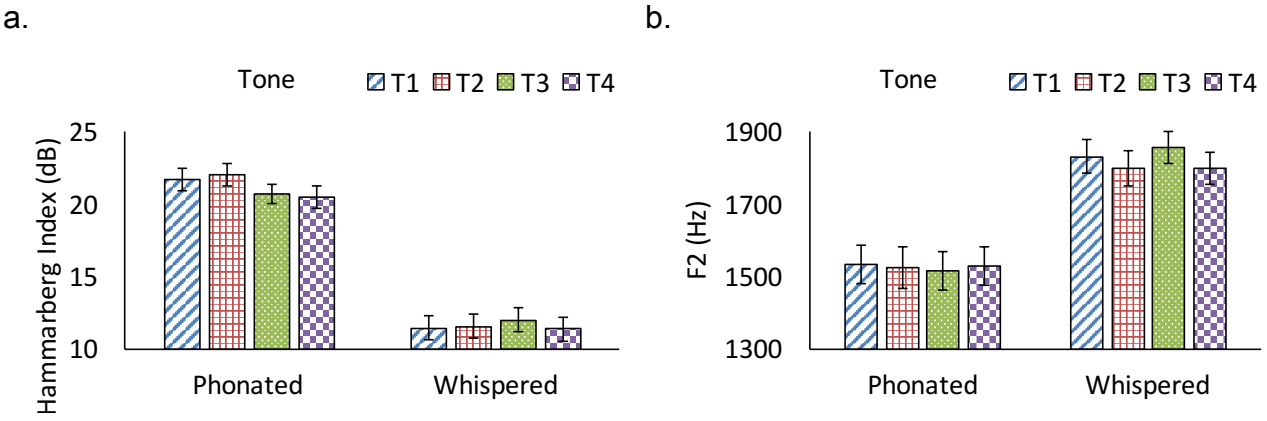


Figure 5: a. Mean values of *Hammarberg Index as a function of phonation type and tone. b. Mean values of F2 as a function of phonation type and tone. The error bars represent standard errors.*

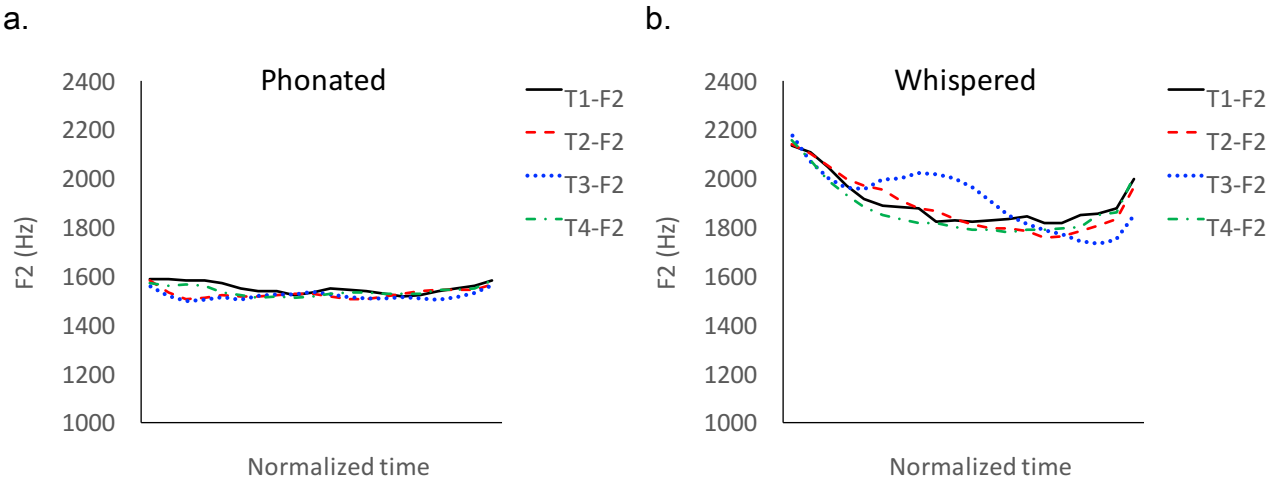


Figure 6: Mean F2 trajectories averaged across all repetitions by all 12 subjects in phonated (a) and whispered (b) utterances.

To sum up this section, of the 9 measurements examined, only F2 shows a possible enhancement for one of the four tones, namely, T3. And even this possibility is questionable as it may be due to errors related to difficulty of tracking formants in whispered utterances.

3.1.3. Effect of Intonation and its interaction with phonation

In Table 4 it can be seen that COG shows both a main effect of intonation (statement vs. question: 861.617 Hz vs. 914.794 Hz) and an interaction between intonation and phonation. Figure 7 shows that in whispers, COG is higher in both intonations than in phonated utterances, and the difference between statement and question is also larger than those in phonated utterances. This could be potentially an enhancement to increase the contrast between question and statement in whispers.

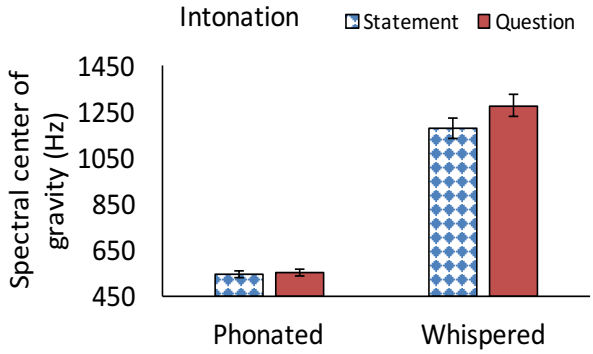


Figure 7: *Mean values of spectral center of gravity (COG) as a function of phonation type and intonation, with standard errors.*

As shown in Table 4, there was no main effect of intonation on Hammarberg Index or Energy below 1000 Hz, another two indicators of spectral tilt, but there were interactions between intonation and phonation on both measurements. Figure 8 shows that intonations in whispers have smaller values of Hammarberg Index (Figure 8a) and Energy below 1000 Hz (Figure 8b), both indicating a flatter spectral tilt than in phonated speech. Furthermore, the spectral slope of question is much flatter than that of statement in whispers. These are therefore further signs of potential enhancement.

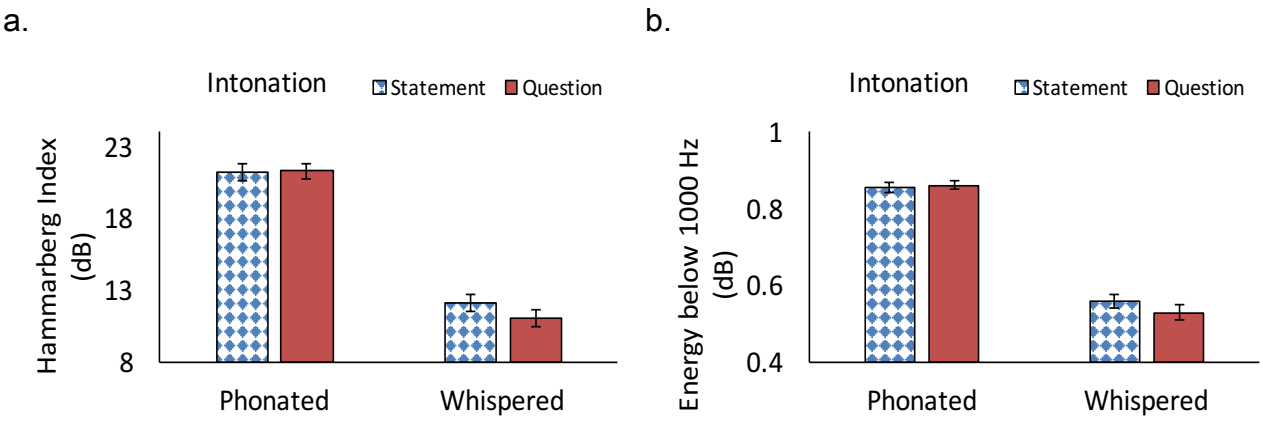


Figure 8: *Mean values of Hammarberg Index (a) and of Energy below 1000 Hz (b) as a function of phonation type and intonation, with standard errors.*

For the rest of measurements, i.e., duration, intensity, F1, F2 and F3, there are no interactions between phonation and intonation. Nor are there higher-order interactions involving three or four factors. In summary, in terms of the effect of intonation, question in whispers has greater COG, but lower Hammarberg Index and Energy below 1000 Hz, all indicating a flatter spectral slope than in statement. These are potential acoustic cues for intonations. But their effectiveness needs to be perceptually assessed.

3.2. Perception of Tone and intonation from Natural Speech Stimuli

All perception results are first normalized to a 0-1 scale, and then analyzed by four-way Repeated Measures ANOVAs, with intonation, phonation, tone and vowel as independent variables. The chance level is 0.25 for tone identification and 0.5 for intonation task.

3.2.1. Perception of Tone

Table 5 shows all significant effects on tone identification in natural utterances. There are main effects of phonation, tone and intonation and their interactions. There are also significant interactions between vowel and other factors, but they will not be discussed because of low relevance to the research questions.

Table 5. *Significant effects by four-way repeated measures ANOVAs on rate of tone identification from natural stimuli.*

Variables	DF	F-Value	P-Value
phonation	1,21	1123.796	<0.0001

tone	3,63	40.478	<0.0001
intonation * tone	3,63	13.692	<0.0001
phonation * tone	3,63	74.135	<0.0001
tone * vowel	6,126	11.092	<0.0001
intonation * phonation * tone	3,63	9.511	<0.0001
intonation * tone * vowel	6,126	10.635	<0.0001
phonation * tone * vowel	6,126	14.083	<0.0001
intonation * phonation * tone * vowel	6,126	5.46	<0.0001

In general, tones are much worse identified in whispered than phonated utterances (49.9% vs. 96.4%). A Student-Newman-Keuls test found significant differences in all tone pairs except T1-T2. Figure 9b shows that all tones were perceived near ceiling in phonated utterances, except T3. In whispers, tone identification dropped drastically, to almost chance level in the case of T1 and T2, but much less for T3 and T4.

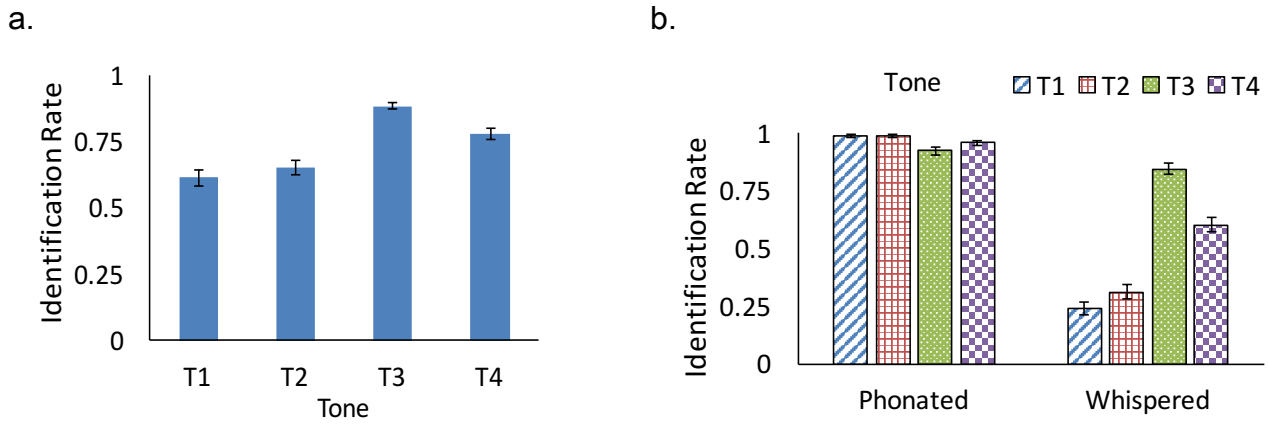


Figure 9. *Rate of tone identification from natural stimuli, as a function of tone (a) and phonation type and tone (b), with standard errors.*

Table 6 is a confusion matrix for tone perception. The numbers in bold correspond to the bars in Figure 9b. As can be seen, there is an overall reduction of tone recognition rate from phonated to whispered tones, indicating that the absence of F_0 has a major impact on

tone perception. Among the whispered tones, T1 is identified more often as T4 than as itself. T2 identification is evenly distributed across all the four tones. T4, at 60.23%, is much better perceived than T1 and T2, although it is sometimes confused with all the other three tones. T3 is much better identified than all the other tones, at a rate as high as 84.47%. These confusion patterns are somewhat similar to those of signal-correlated noise found in Author and Author (1992). As found in that study and also seen in Figure 3, the likely cues are the long duration and bimodal intensity profiles of T3. As will be seen in Section 3.3, similarly high identification rate of T3 is also found in phonated utterances when F_0 was removed.

Table 6. *Confusion Matrix of Tone Identification Rates in Natural Stimuli.*

	Heard / Original	T1(%)	T2(%)	T3(%)	T4(%)
Phonated	T1	**98.48**	0.76	0.38	0.38
	T2	0.76	**98.86**	0.00	0.38
	T3	0.00	7.20	**92.42**	0.38
	T4	0.00	0.38	3.79	**95.83**
Whispered	T1	**23.86**	20.45	8.33	47.35
	T2	20.83	**31.06**	27.27	20.83
	T3	0.76	12.50	**84.47**	2.27
	T4	19.32	12.12	8.33	**60.23**

3.2.2. Perception of Intonation

As shown in Table 7, there are main effects of intonation, phonation and their interaction. Generally, more statements were identified than questions (0.851 vs. 0.547). As seen in Figure 10, the gap between the identification rates for statements and questions becomes much wider in whispers, to the extent that questions are recognized below chance, as can be also seen in the confusion matrix in Table 8. Note that the greater identification rate of statements does not necessarily mean that they were more correctly recognized. Rather, statement is likely treated as a default choice when identification was impossible. This can be tested in future studies by adding “cannot decide” as a choice for answer.

Table 7. *Significant effects by four-way repeated measures ANOVAs on rate of intonation identification from natural stimuli.*

Variables	DF	F-Value	P-Value
intonation	1,21	26.835	<0.0001

phonation	1,21	66.215	<0.0001
tone	3,63	3.928	0.0124
intonation * phonation	1,21	45.504	<0.0001
intonation * tone	3,63	40.189	<0.0001
intonation * vowel	2,42	4.768	0.0136
phonation * tone	3,63	11.498	<0.0001
phonation * vowel	2,42	15.303	<0.0001
tone * vowel	6,126	3.062	0.0079
intonation * phonation * tone	3,63	18.59	<0.0001
intonation * phonation * vowel	2,42	14.326	<0.0001
intonation * tone * vowel	6,126	7.266	<0.0001

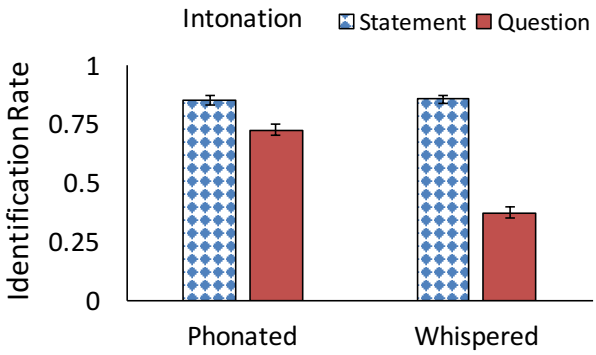


Figure 10: *Rate of identification of intonation from natural stimuli as a function of phonation type and intonation, with standard errors.*

Table 8. *Confusion matrix of intonation identification from natural stimuli.*

	Heard / Original	Statement (%)	Question (%)
Phonated	Statement	**84.85**	15.15
	Question	27.65	**72.35**
Whispered	Statement	**85.42**	14.58
	Question	62.88	**37.12**

3.3. Perception of Tone from Amplitude-modulated Noise

As described in 2.2.2, amplitude-modulated noise was generated based on both phonated and whispered utterances, and was used as stimuli in a tone identification experiment. The results are shown in Table 9. There is a marginal effect of phonation, a strong effect of tone, and a strong interaction between the two factors.

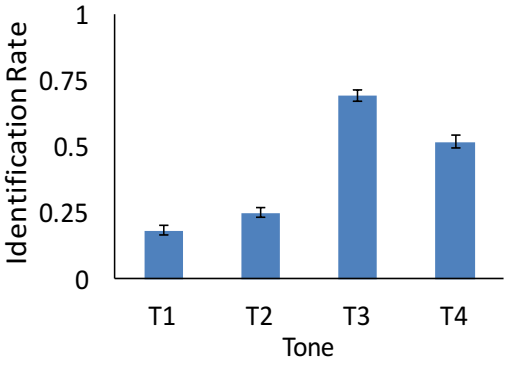
The phonated tones were less well identified than whispered tones (0.389 vs. 0.432). Figure 11a shows overall identification of the four tones in both phonated and whispered utterances. As can be seen, T3 is still the best judged, and T4 is the second best, while T1 and T2 are around chance. Figure 11b shows identification rates in the two phonation types separately. Also Table 10 shows confusion matrices for the two phonation types, respectively. As can be seen, there is not much difference between the two conditions, but tonal identification is slightly better in whispered utterances than in phonated utterances. Upon further examination as to whether this is due to possible enhancement for whispered tones, we discovered the slightly worse identification rate in phonated utterances was mainly due to relatively poorer performance in question intonation, as shown in Figure 12. For unclear reasons, much of the non-F_0 cues in the phonated utterance that can be retained in the amplitude-modulated noise was rendered absent by question intonation. In particular, in statements, T3 was well identified at 79%, compared to 48% in questions. In summary, therefore, the perception results for amplitude-modulated noise do not seem to have provided evidence of enhancement for whispered tones.

Table 9. *Repeated Measures ANOVAs of Significant Tone Identification Rates in Synthesized Stimuli.*

Variables	DF	F-Value	P-Value
phonation	1,21	5.268	0.0321
tone	3,63	74.348	<0.0001
vowel	2,42	4.646	0.015
intonation * tone	3,63	11.392	<0.0001
intonation * vowel	2,42	3.828	0.0297
phonation * tone	3,63	18.358	<0.0001

tone * vowel	6,126	6.554	<0.0001
intonation * phonation * tone	3,63	7.514	0.0002
phonation * tone * vowel	6,126	3.392	0.0039

a.



b.

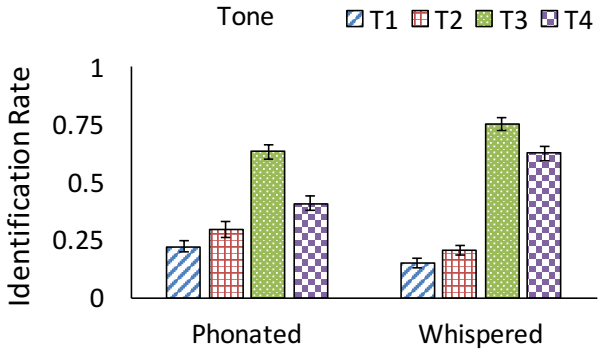


Figure 11: *Rate of tone identification from amplitude-modulated noise as a function of tone (a) and phonation type and tone (b), with standard errors.*

Table 10. *Confusion Matrix of Tone Identification Rates in Synthesized Stimuli.*

	Heard / Original	T1(%)	T2(%)	T3(%)	T4(%)
Phonated	T1	**22.00**	26.50	10.20	41.30
	T2	18.90	**29.50**	21.20	30.30
	T3	6.80	18.60	**63.30**	11.40
	T4	22.00	22.30	14.80	**40.90**
Whispered	T1	**14.77**	22.73	19.70	42.80
	T2	14.00	**20.50**	23.90	41.70
	T3	1.90	15.50	**75.00**	7.60
	T4	12.50	13.60	11.40	**62.50**

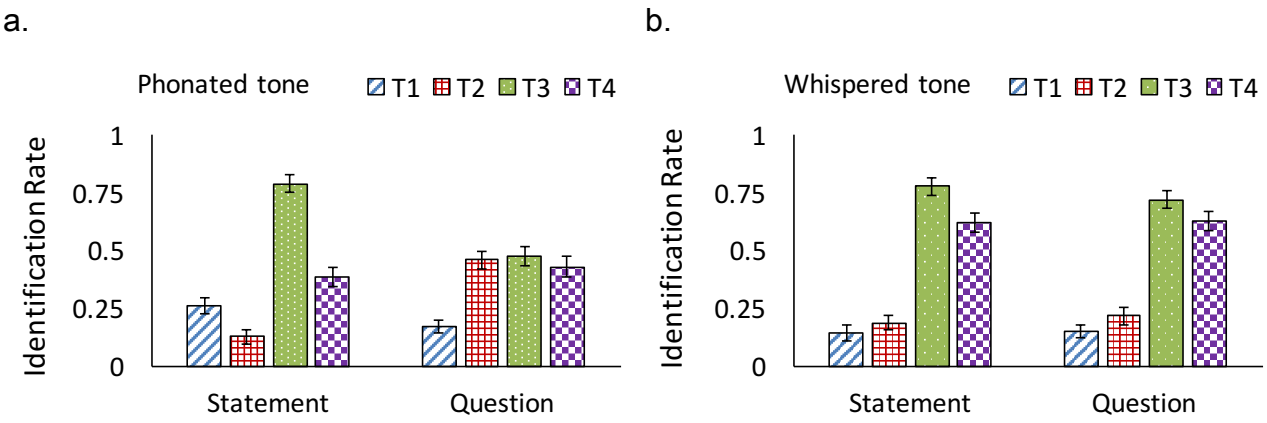


Figure 12: *Rate of tone identification from amplitude-modulated noise based on phonated (left) and whispered tones (right) as a function of tone and intonation, with standard errors.*

4. Discussion and Conclusion

The aim of the present study is to assess whether Mandarin has developed special acoustic cues, or *production enhancement*, to aid the perception of tone and intonation. One production and three perception experiments were conducted to determine if there are special articulatory manoeuvres in whispers that are either absent in phonated speech or are exaggerations of what is already there. The production experiment examined nine measurements of duration, intensity spectral tilt and formants. For tone, none of these measurements showed evidence of enhancement, with only the possible exception of F2. Upon closer examination, however, even the special behaviour of F2 in whispers was determined to be likely due to formant tracking errors rather than genuine enhancement. For intonation, it was found that question in whispers has greater COG, but lower Hammarberg Index and Energy below 1000 Hz, all indicating a flatter spectral slope than in statement.

The three perception experiments examined whether any of the potential acoustic cues or those not detected in the measurements were used in the identification of tone and intonation. For tone, the results can be best seen by comparing Figures 9b and 11b. Figure 9b shows that when F_0 is absent in whispers, tone identification rate dropped overall, yet T3 and T4 were still recognized well above chance. Figure 11b shows that, however, when F_0

is removed from phonated utterances in amplitude-modulated noise, the patterns of tone recognition became very similar to those of whispered utterances. This indicates that the tone-specific non-F_0 acoustic patterns were already in the phonated speech. In particular, the high recognition rate of T3 is likely due to its extra-long duration (Figure 1 and related discussion), They are just not very useful given the dominance of F_0 when it is present (Abramson, 1972; Lin, 1988). Furthermore, although the tone identification was slightly better from amplitude-modulated noise when the originals were whispers than when they were phonated utterances (Table 9 and Figure 11b), the difference seems to be mainly due to changes in the non-F_0 cues of phonated speech as shown in Figure 12a. It is hard to imagine that speakers have somehow developed ways to prevent the similar changes in the non-F_0 cues when they whisper so as to preserved tone perception in questions. Rather, it is likely that the changes are concomitant with the F_0 changes in phonated intonation which is absent in whispers due to lack of voicing.

As a further note, even the better-than-chance tone perception for some of the tones may largely disappear in whispered continuous speech, as found by Gu *et al.* (2016) for Singapore Mandarin. One likely reason is that at least the duration cues would become rather ineffective in continuous speech, as all the four Mandarin tones become very similar in duration (Author, 1997). Furthermore, the shortening of T3 is accompanied by the loss of the final rise often found in isolation, which would also lead to the loss of the dipping amplitude profile found in this tone (Author and Author, 1992).

With regard to the perception of intonation, the flattened spectral slope as indicated by increased COG and decreased Hammarberg Index and Energy below 1000 Hz, did not seem to help the identification of questions. As shown in Figure 10, questions in whispers were much less recognized than statements. Although this could be explained as partially due to a bias toward the latter in case of ambiguity, there is no evidence of enhancement for intonation perception afforded by the spectral flattening.

The present data therefore have provided virtually no evidence of effective production enhancement for either tone or intonation in whispered Mandarin. This finding seems to be at odds not only with previous findings about whispered speech in Mandarin (Gao, 2002; Li and Guo, 2012; Liu and Samuel, 2004) and other languages (Heeren and Van Heuven, 2014; Żygis *et al.*, 2017), but also with the widely recognised importance of both tone and

intonation in phonated speech (Bolinger, 1983; Chao, 1968; Hirst and Di Cristo, 1998). As shown by Surendran and Levow (2004), the functional load of tone is as high as that of vowel in Mandarin. Why, then, hasn't Mandarin developed effective enhancements to aid the perception of tone and intonation to make up for the absence of F_0 in whispers? One possibility is that the functional load of tone and intonation is not as high as it is usually understood. In Surendran and Levow (2004), it is shown that the functional load of tone based on word is one order of magnitude smaller than that based on syllable. This is because words, being disyllabic in Mandarin on average, are longer than syllables and hence less likely to be homophonic. Zhang *et al*. (2010) further show that the functional load based on sentence is a further order of magnitude smaller than that based on word. These findings provide strong evidence of a high level of redundancy in speech. Such redundancy would mean that there is unlikely to be a pressing need to develop enhancement strategies just to make up for the absence of F_0 in whispers. Furthermore, though there is not yet formal research to our knowledge, it is likely that whispering typically occurs in situations where people know each other well and the topic of conversation is familiar to the participants. This would further reduce the incentive for developing production enhancement strategies that need to be not only shared between the current conversation participants, but also preserved and reused in future occasions when whispering is needed.

Acknowledgements

We thank the Chinese Scholarship Council for supporting the first author for her one-year study in Faculty of Linguistics, Philology and Phonetics at University of Oxford. Part of this project was presented in *Speech Prosody 2016* under the support by the International Exchange Program for Graduate Students to the first author. Our thanks also go to Professor John Coleman in Oxford, Professor Qiuwu Ma in Fudan University, Professor Daniel Hirst, Professor Jie Liang and Dr. Ting Wang at Tongji University, Professor Hongwei Ding at Shanghai Jiao Tong University and Dr. Marjoleine Sloos from Koninklijke Nederlandse Akademie van Wetenschappen. We also owe many thanks to participants from both Tongji and Oxford sites.

References

Abramson, A. S., 1972. Tonal experiments with whispered Thai. In Valdman, A. (Ed.), Papers on linguistics and phonetics in memory of Pierre Delattre. The Hague: Mouton, pp. 31-44.

Boersma, P., 2001. Praat, a system for doing phonetics by computer. Glot International, 5, pp. 341-345.

Bolinger, D., 1982. Intonation and Its Parts. Language 58 (3), pp.505-533.

Chang, C. and Yao, Y., 2007. Tone production in whispered Mandarin. Proceedings of the 16th International Congress of Phonetic Sciences, August 6-10, Saarbrücken, Germany, pp. 1085-1088.

Chao, Y. R., 1968 A Grammar of Spoken Chinese. Berkeley, CA: University of California Press.

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

Hammarberg, B., Fritzell, B., Gaufin, J., Sundberg, J. and Wedin, L., 1980. Perceptual and Acoustic Correlates of Abnormal Voice Qualities. Acta Oto-Laryngologica, 90(1-6), pp. 441-451.

Heeren, W. F. L. and Van Heuven, V. J., 2014. The interaction of lexical and phrasal prosody in whispered speech. The Journal of The Acoustical Society of America, 136 (6), pp. 3272-3289.

Heeren, W. F. L., 2015. Vocalic correlates of pitch in whispered versus normal speech. The Journal of The Acoustical Society of America, 138(6), pp. 3800-3810.

Higashikawa, M. and Minifie, F. D., 1999. Acoustical-Perceptual Correlates of "Whisper Pitch" in Synthetically Generated Vowels. Journal of Speech, Language, and Hearing Research, 42, pp. 583-591.

Hirst, D. and Di Cristo, A., 1998. Intonation Systems -- A survey of twenty languages: Cambridge University Press.

Hockett, C. F., 1955. A Manual of Phonology. Baltimore, Waverly Press.

Hockett, C.F., 1967. The Quantification of Functional Load. Word, 23, pp. 320-339.

Honda, K., Hirai, H., Masaki, S. and Shimada, Y., 1999. Role of Vertical Larynx Movement and Cervical Lordosis in F0 Control. Language and Speech, 42(4), pp. 401-411.

Jiao, L., Ma, Q.-W., Wang, T. and Xu, Y., 2015. Perceptual cues of whispered tones: Are they really special? Proceedings of Interspeech 2015, September 6-10, Dresden, Germany, pp.2361-2365.

Jiao, L. and Xu, Y., 2016. Interactions of tone and intonation in whispered Mandarin. Proceedings of Speech Prosody 2016, May 31-June 3, Boston, USA, pp.94-98.

Laver, J., 1994. Principles of Phonetics. Cambridge: Cambridge University Press.

Li, B. and Guo, Y.-M., 2012. Mandarin tone contrast in whisper. In Proceedings of Tonal Aspects of Languages-Third International Symposium.

Lin, M.-C., 1988. Putonghua shengdiao de shengxue texing he zhijue zhengzhao [The acoustic characteristics and perceptual cues of tones in Standard Chinese]. Zhongguo Yuwen [Chinese Linguistics], 204 (3), pp. 182-193.

Liu, S.-Y. and Samuel, A. G., 2004. Perception of Mandarin Lexical Tones when F0 Information is Neutralized. Language And Speech, 47 (2), pp. 109-138.

Meyer-Eppler, W., 1957. Realization of Prosodic Features in Whispered Speech. The Journal of The Acoustical Society of America, 29(1), pp.104-106.

Moisik, S. R., Lin, H. and Esling, J. H., 2014. A study of laryngeal gestures in Mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (SLLUS). Journal of the International Phonetic Association, **44**(1), pp. 21-58.

Shannon, C. E., 1948. A Mathematical Theory of Communication. Bell System Technical Journal, 27(4), pp. 623-656.

Surendran, D. and Niyogi, P., 2006. Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals. Chapter in Competing Models of Linguistic Change: Evolution and Beyond. In commemoration of Eugenio Coseriu (1921-2002). Ole Nedergaard Thomsen (ed), Amsterdam & Philadelphia: Benjamins.

Surendran, D. and Levow, A., 2004. The Functional Load of Tone in Mandarin is as High as that of Vowels. In Proceedings of the International Conference on Speech Prosody 2004, pp. 99-102, Nara, Japan.

Tartter, V. C., 1989. What's in a whisper? The Journal of The Acoustical Society of America, 86(5), pp. 1678-1683.

Thomas, I. B., 1969. Perceived Pitch of Whispered Vowels. The Journal of The Acoustical Society of America, 46 (2B), pp. 468-470.

Wang, W. S.-Y., 1967. The Measurement of Functional Load. Phonetica, 16, pp. 36-54.

Whalen, D. H. and Xu, Y., 1992. Information for Mandarin Tones in the Amplitude Contour and in Brief Segments. Phonetica, 49 (1), pp. 25-47.

Wise, C. M. and Chong, L.-P., 1957. Intelligibility of Whispering in A Tone Language. Journal of Speech and Hearing Disorders, 22 (3), pp. 335-338.

Xu, C.-X. and Xu, Y., 2003. Effects of consonant aspiration on Mandarin tones. Journal of the International Phonetic Association, 33 (2), pp. 165-181.

Xu, Y., 2004. Understanding tone from the perspective of production and perception. Language and Linguistics, 5, pp. 757-797.

Xu, Y., 2007-2015. FormantPro. praat. Retrieved 19 August 2017, from http://www.phon.ucl.ac.uk/home/yi/FormantPro/

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

Yip, M., 2002. Tone. Cambridge: Cambridge University Press.

Zemlin, W. R., 1988. Speech and Hearing Science — Anatomy and Physiology. Englewood Cliffs, New Jersey: Prentice Hall.

Zhang, J.-S., Li, W., Hou, Y.-X., Cao, W. and Xiong, Z.-Y., 2010. A study on functional loads of phonetic contrasts under context based on mutual information of Chinese text and phonemes. Proceedings of Chinese Spoken Language Processing (ISCSLP), 7th International Symposium on, pp. 194-198.

Żygis, M., Pape, D., Koenig, L., Jaskuła, M. and Jesus, L., 2017. Segmental cues to intonation of statements and polar questions in whispered, semi-whispered and normal speech modes. Journal of Phonetics, 63, pp. 53-74.