Perceptual Cues of Whispered Tones: Are They Really Special?

Li Jiao

School of Foreign Languages, Tongji University, Shanghai, China

2013li_jiao@tongji.edu.cn

Abstract

The purpose of this dissertation is to find out whether Mandarin Chinese has developed effective strategies to convey tonal information in whispered speech. Three research questions are asked: 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?

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 F0, 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 [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 F0 is absent in signal-correlated noise [7], but their contribution to tone perception is small (about 3%) when F0 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 F0 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 tone without affecting communication in a relatively ideal listening environment. Such redundancy is needed in adverse conditions such as at cocktail parties [11].

The whole project consists of three major experiments. In the first experiment, I recorded phonated and whispered tones in monosyllabic words, and analyzed the acoustic properties of the tonal contrast. The database for each speaker is composed of five sets of syllable with vowel glide onset (/a/, /i/, /u/ and /y/) and three sets with obstruct onsets (cha/, ce/ and /ch/). They carry the four lexical tones of Mandarin: Tone 1 (T1), Tone 2 (T2), Tone 3 (T3) and Tone 4 (T4). These target words were recorded either with or without a carrier (‘zhè gè zì nián ...’ ‘The word is read as ...’), and in three modes: citation, interactive statement, and interactive question. Two native speakers of Mandarin (a male and a female) took part in the recording. 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.

The acoustic analysis was done with a modified version of ProsodyPro, a Praat script for large-scale prosody analysis [12]. With the script, I 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 [13]
- 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.

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.

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. To see if there are more detailed tone-specific intensity patterns, the mean time-normalized temporal profiles of intensity of the four tones in phonated as well as whispered speech showed that 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 I 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.

Overall, the initial acoustic findings from the female speaker showed that there are acoustic properties corresponding to the four tones of Mandarin other than F0, 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. The limitation at the current stage is lack of speakers. I have recorded ten more speakers (five males and five females) and will perform the same analyses. Then I will check whether the initial findings are speaker-specific or universally meaningful.

In the second experiment, the phonated and whispered syllables /xi, yi and /yi/ by the female speaker were used as perception stimuli. Tonal intelligibility was examined through two conditions: 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 [14] and imposed onto pink noise of the same duration. The pink noise was generated by filtering white noise in Praat [14], the Tier of each syllable was extracted in Praat [14] and imposed onto pink noise of the same duration. The resulting signal therefore contained both the duration and amplitude undulation of the syllable, but no spectral information.

Twenty native Mandarin listeners (ten males and ten females) living in China participated as subjects. 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.

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.

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.

For the amplitude-modulated noise, 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. This advantage cannot be seen in the original whispered T4, however. This suggests that it could have been an artifact of the generation of the amplitude-modulated noise.

Additionally, confusion matrices for the perception of the original and amplitude-modulated tones showed that phonated tones were all perceived accurately, except that T3 and T4 had somewhat lower accuracy. In whispered speech, in contrast, T3 is the best perceived. This could be due to its longer duration, relatively low intensity, bi-modal intensity profile, and greater spectral tilt. T1 and T2 were poorly perceived. T1 was heavily confused with both T4 and T2, while T2 is confused with all the other tones. Correspondingly, 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.

To sum up the first two experiments, although I have found consistent acoustic cues that may have led to the good recognition of some of the Mandarin tones whispered in isolation, I 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.

The further plans for completing my PhD dissertation are: 1) additional work for the first experiment: analyze more speakers in terms of acoustic cues in order to obtain evidence to verify my current conclusion; 2) start to run the third experiment (it will consist of several sub experiments): using an articulatory synthesizer to synthesize whispered tones and examine their perceptual effectiveness.

The purpose of the third experiment is to further test the acoustic and perceptual cues in whisper-like utterances, trying to answer: 1) Is it be possible to synthesize whispered Mandarin tones that can be identified as well as the human-produced ones? 2) Is it possible to find ways to enhance synthetic whispered Mandarin tones so that they are perceived better than the human whispered ones? Answering these questions, especially the second one, may help to show that the lack of special cues in natural whispers is not due to lack of means to do so, but probably due to lack of need for it. I have not worked out the full details of the experiment, but I would try VocalTractLab, a three-dimensional model of the human vocal tract [15], which recently included a highly effective glottal model [16]. With the supporting acoustic data from the first experiment, I will modulate the basic parameters and get synthesized sounds. Then I will run a subsequent perception experiment to test them.

In summing up, 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 [2,17]. Second, acoustic analysis showed a number of possible acoustic cues that varied with tone, including, duration, intensity, temporal profile of intensity and
mechanism of whispered tones via the synthesized sounds. In addition, it understanding of “special” cues to whispered Mandarin tones spectral tilt in whispered tones show similar patterns. Mandarin tones in phonated speech, and the variation of helps the perception of whispered boundary tones in Dutch. In example, has shown that artificially increased spectral tilt enhancement has been demonstrated to be possible. [6], for tonal cues, as claimed by some early studies [2]. to have developed special strategies to enhance the secondary from the data of one female speaker, Mandarin does not seem interacti of their enhancement in whispering, as there is lack of consistent cross 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. In the current results, however, spectral tilt already varies with Mandarin tones in phonated speech, and the variation of spectral tilt in whispered tones show similar patterns.

This PhD dissertation will then contribute to the new understanding of “special” cues to whispered Mandarin tones. In addition, it will provide more evidence to the articulation mechanism of whispered tones via the synthesized sounds.

References