Perception of Word Tones in Mandarin Whiskered Speech

John Kwok-Ping Tse*

ABSTRACT

The recognition of word tones in whispered speech may reveal the significance of fundamental frequency variation in speech sound perception. However, relatively few studies have been done on this aspect of tone perception and many issues remain unresolved. Among them are the question of how well whispered word tones can be recognized and the question of whether or not some tones are perceptually more salient than others. The present study attempts to address these two questions through a perception experiment. Our findings suggest that when fundamental frequency variation is absent, as in whispered speech, perception of word tones is greatly hampered if proper and larger linguistic contexts are also excluded, suggesting that not much word tone is transmitted in whispered speech. This is contradictory to Jensen’s (1958) finding and provides some supporting evidence to Miller’s (1961) finding, thus bringing some additional evidence to the answer and solution of the first question. As for the second question, we have significant findings which show that Tone 2 and Tone 3 in Mandarin are perceptually more salient in whispered speech; but we fail to find any acoustic correlates in the test items to account for this aspect of saliency. In the light of these results, we suggest that further studies be done to explore the possibility of the existence of other kinds of acoustic correlates and replications of the present experiment be carried out in order to seek better explanations for the perceptual saliency of Tone 2 and Tone 3 in Mandarin whispered speech.

Key Words: whisper, fundamental frequency, perceptual saliency

*Professor, Department of English, National Taiwan Normal University
I. INTRODUCTION

The study of the perception of speech has always been a major concern in acoustic phonetics and speech science in general. Speech scientists, psychologists, and psychoacoustics researchers have long tried to studied speech perception by controlling and thus isolating certain acoustic properties in the speech signals in various perception/recognition experiments. The major idea is to explore what roles these acoustic properties will play in the process of perceiving speech sounds. We of course understand that in ordinary speech, the acoustic signals are usually kept intact, with all concomitant information present in the signals. The end results of these controlled tests, therefore, give us information about speech sound perception in more or less "degraded" listening conditions in the sense that the test signals are somehow modified, filtered, or masked. It is of interest to point out that "many of the advances in our understanding of normal speech have come from the study of speech perception in degraded listening conditions." (Tartter 1989, p.1678). However, as a general form of degraded speech, the whisper has not been a frequent subject of inquiry in the literature (Tartter 1989). Fewer still are studies on the perception/recognition of tones in whispered speech in general and in Mandarin whispered speech in particular.

On the segmental level, studies have been carried out on the formant frequencies and amplitudes of phonated and whispered vowels (Kallail and Emanuel 1984, 1985), voicing in whispered consonants (Dannenbring 1980, as reviewed in Tartter 1989), bilabial closure durations (Schwartz 1972), place and manner of articulation and voicing in whispered consonants (Tartter 1989), etc.

On the prosodic level, studies have been concentrated on the perception of whispered tones and pitch, since it is of interest to study the difficulty involved because of the absence of periodic vibration of the vocal cords (namely, fundamental frequency by definition) in whispered speech, particularly under experiment conditions where context offers no help

Wise and Chong (1957), Jensen (1958) and Miller (1961) were among the earliest studies on recognition of word tones. The recognition score obtained by Wise and Chong for Mandarin is 62% (as reviewed in Jensen 1957, p. 188). The range of recognition scores obtained by Jensen in his tests are: Norwegian--53-73% (6 tests), Swedish--100% (1 test), Slovenian--71-85% (5 tests), Mandarin--73-88% (4 tests). All Jensen's test scores are thus well above chance level. Miller, on the other hand, replicating Jensen's procedures, obtained only a range of 22-58% correct recognition for Vietnamese. While Jensen could claim that word tones are reflected somehow in whispered speech, Miller (1961, p.15) could only say that "very little word tone is transmitted in Vietnamese whispered speech." The discrepancy for word tone recognition in whispered speech is rather great and it has remained unresolved over the years. Possible reasons for this discrepancy might be the relatively small number of subjects involved in these early studies and the relatively informal methods used in them.

The main purpose of the present study is to conduct a more formal perception test involving more subjects so as to seek a better understanding of the recognition of Mandarin tones in whispered speech. We are not only interested in probing into the accuracy level of recognition so that we may shed some light on the above-mentioned discrepancy, but we are also interested in looking at other aspects of tone perception in Mandarin in the absence of fundamental frequency and context.

Specifically, we are interested in the following questions:
(1) What is the average level of accuracy for perception of whispered tones in Mandarin in minimal contrast contexts?
(2) Is there any difference in the accuracy level for perception of different tones in Mandarin whispered speech? In other words, is it possible that some tone(s) is more perceptually salient than others in Mandarin whispered speech?
(3) Can any acoustic correlates be found if the answer to (2) is positive?

(4) Is there any gender difference in the perception of whispered tones in Mandarin?

II. METHOD

A. Speech Material

Four sets of bisyllabic phrases involving minimal contrast in the four tones of Mandarin were designed, with each set containing 4 phrases, making a sum total of 16 test items. These 16 items are listed in Table 1 below.

<table>
<thead>
<tr>
<th>Tone 1</th>
<th>Tone 2</th>
<th>Tone 3</th>
<th>Tone 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) 大威 [wei ˥]</td>
<td>大維 [wei ˩]</td>
<td>大偉 [wei ˩]</td>
<td>大衛 [wei ˥]</td>
</tr>
<tr>
<td>(b) 大飛 [fei ˥]</td>
<td>大肥 [fei ˩]</td>
<td>大匪 [fei ˩]</td>
<td>大肺 [fei ˧]</td>
</tr>
<tr>
<td>(c) 大巴 [ba ˥]</td>
<td>大拔 [ba ˩]</td>
<td>大把 [ba ˩]</td>
<td>大壩 [ba ˧]</td>
</tr>
<tr>
<td>(d) 大媽 [ma ˥]</td>
<td>大麻 [ma ˩]</td>
<td>大馬 [ma ˩]</td>
<td>大罵 [ma ˧]</td>
</tr>
</tbody>
</table>

The items in (a) within Table 1 were placed in a carrier sentence "他的名字叫____" (His name is____). The carrier sentence for items in (b) was "他的綽號叫____" (His nickname is____) and the items in both (c) and (d) were placed in the carrier sentence "這兩個字唸____" (These two words are read as____). Therefore, although our attention was focused on the contrast of the four tones (e.g., 威 [wei ˥], 維 [wei ˩], 偉 [wei ˩], 衛 [wei ˧], etc.), the actual whispered speech presented to the subjects for perception were whole sentences with the test items being placed in the carriers (e.g. 他的名字叫大威, 他的綽號叫大肥, 這兩個字唸大把, 這兩個字唸大罵, etc.).

Eight other items with similar carrier sentences were also designed and included in the perception test to serve as distractors. See the Appendix for a complete listing of all the 24 items (16 test items and 8 distractors) with their carriers.

B. The Subjects
Seventy-seven young college students served as our subjects. Among them, 17 are male students and 60 are female students. Their age ranged from 18 to 30. There was a variety of academic majors among the subjects. The majority of them, however, were language majors and science majors. Care was taken to ensure that none of our subjects had any problems in hearing.

C. Procedure

The test items and the distractors occurring in the carriers were randomized and then pronounced in whisper by an adult male speaker of Mandarin who is accurate in pronunciation. This whispered version of all 24 sentences was recorded on tape in a sound-proof recording room by a high quality cassette tape recorder. The speaker was asked not to whisper loud into the microphone but to whisper the sentences in as natural and comfortable a manner as possible so that no particular and deliberate increase in syllable amplitude was made. This is very important since increasing syllable amplitude in whispered speech often results in some degree of voicing. A short pilot test was then administered to a small group of students to find out whether the tape-recorded test sentences could be heard clearly in quiet rooms and the result showed that they could. The tape was played back to the subjects in the perception/listening test.

The test was administered in sound-proof rooms to the subjects in several separate sessions. Before the subjects listened to the recording, they were given an answer sheet with the four choices of the test items given for each of the 24 test sentences. The choices are minimal word sets in terms of the four tones. The subjects were told that after they listened to each sentence they were to make a choice as to which of the four choices was the one they thought they heard. For example, after they heard the sentence 他的名字叫大偉 (His name is Da-wei), the subjects were to choose which one of the four choices (大威, 大維, 大偉, 大衡) they thought they heard. Each sentence was played back twice and a 6-second pause was inserted between two sentences.
After the test, the results were scored, tabulated and then submitted to statistical analysis. At the same time, the 16 recorded test items were spectrographically analyzed by using the KAY 5500 DSP Sonograph to determine the duration, the overall intensity, and the formant frequencies of the contrastive syllables of the test items (e.g., 威, 維, 偉, 衛).

III. RESULTS

Each correct response in the perception test was given 1 point and with 16 test items (4 for each tone), the maximum total scores a subject could get was 16. The average total scores obtained by the subjects as a group is 8.61. This score represents 54% out of a total of 16, which shows that about half of their judgements on the type of tones in whispered speech were wrong.

There is no significant gender difference in the subjects' perception capability. While male subjects scored an average of 8.52 (N=17, SD=2.2), female subjects did only slightly better, with an average score of 8.63 (N=60, SD=2.0). A comparison of these two average scores shows that the difference is not significant (t=0.18, df=75, N.S).

Since the range of the age of the subjects is small, 18 to 30, we use the age 20 as a reference point and divide them into two groups. Those subjects who were 20 and below scored an average of 8.41 (N=27, SD=2.1) and those who were above 20 obtained an average score of 8.72 (N=50, SD=2.0). But the difference is not significant (t=0.64, df=75, N.S.).

A confusion matrix is constructed to show how each one of the four tones were perceived by the subjects and when their perception was wrong, how their wrong judgements were distributed. For each tone, there are 308 responses (4 test items for each tone judged by 77 subjects). Table 2 summarized the distribution of these responses.
Table 2. Confusion Matrix of the Four Tones

<table>
<thead>
<tr>
<th></th>
<th>RESPONSES (Perceived as)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tone 1</td>
</tr>
<tr>
<td>Tone 1</td>
<td>71</td>
</tr>
<tr>
<td>Tone 2</td>
<td>10</td>
</tr>
<tr>
<td>Tone 3</td>
<td>33</td>
</tr>
<tr>
<td>Tone 4</td>
<td>34</td>
</tr>
<tr>
<td><strong>Column Total</strong></td>
<td>148</td>
</tr>
</tbody>
</table>

The chi-square test shows that the distribution in Table 2 is significant ($X^2 = 660.5$, df=9, P<0.001). A careful look at the distribution reveals that generally when confusion happened, more whispered tones were wrongly perceived as Tone 2 and Tone 3 than as Tone 1 and Tone 4. The column totals, which include both correct and wrong responses, also reflect the fact that proportionally there are more responses clustered around Tone 2 and Tone 3, suggesting that these two tones might be perceptually more salient in whispered speech.

When one occurrence of a wrong judgment is given one point, the average scores of these wrong judgements are calculated and summarized for each tone in Table 3 below.

Table 3. Average Scores of Wrong Responses

<table>
<thead>
<tr>
<th>Items Wrongly perceived as</th>
<th>Mean Scores of Wrong Responses</th>
<th>N</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tone 1</td>
<td>1.0</td>
<td>77</td>
<td>0.89</td>
</tr>
<tr>
<td>Tone 2</td>
<td>2.9</td>
<td>77</td>
<td>1.59</td>
</tr>
<tr>
<td>Tone 3</td>
<td>2.5</td>
<td>77</td>
<td>1.37</td>
</tr>
<tr>
<td>Tone 4</td>
<td>1.1</td>
<td>77</td>
<td>0.95</td>
</tr>
</tbody>
</table>

A oneway ANOVA is performed and the results show that the variance of the subjects wrong judgements are significant [$F=42.79$, df=(3, 304), P<0.001]. The Tukey test for a
posterori comparison between the means is calculated and four significant contrasts are shown: Tone 1 vs Tone 2; Tone 1 vs Tone 3; Tone 4 vs Tone 2; Tone 4 vs Tone 3. The comparisons of Tone 2 vs Tone 3 and Tone 1 vs Tone 4 are not significant. This reflects essentially the same findings in the confusion matrix above; namely, when errors of perception occurred, the stimuli tended to be misjudged as Tone 2 and Tone 3 more often than as Tone 1 and Tone 4. But the ANOVA also shows that though Tone 2 and Tone 3 may be perceptually more salient, the difference between their possible saliency is not significant.

In order to explore the possible relationship between the differences in the perception of the four tones and the acoustic properties of these tones, we further analyze the average duration, overall intensity, and formant frequencies of the 16 contrastive syllables in the test stimuli (4 for each tone). These scores are summarized in Table 4 (for duration and intensity) and Table 5 (for formant frequencies) below. While measurements of duration and formant frequencies are straightforward, the overall intensity is obtained by calculating the average dB values at the initial point, the peak and the end point of the intensity curves of the contrastive syllables. Furthermore, when formant frequencies are examined, the 16 syllables are divided into two groups, those with a nuclear vowel [a] (巴, 拔, 把, 場, 媽, 麻, 馬, 罵) and those with a nuclear vowel [e] (威, 維, 偉, 衛, 飛, 肥, 匪, 肺). Measurements are made with respect to the lower three formants (F1, F2, F3) of the nuclear vowels,
Table 4. Average Duration, Intensity of the Contrastive Syllables

<table>
<thead>
<tr>
<th>Tone</th>
<th>Average Duration (in msec)</th>
<th>Average Intensity (in dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tone 1</td>
<td>320</td>
<td>68</td>
</tr>
<tr>
<td>Tone 2</td>
<td>377</td>
<td>62</td>
</tr>
<tr>
<td>Tone 3</td>
<td>325</td>
<td>66</td>
</tr>
<tr>
<td>Tone 4</td>
<td>270</td>
<td>68</td>
</tr>
</tbody>
</table>

Table 5. Average Formant Frequencies of the Contrastive Syllables

<table>
<thead>
<tr>
<th>Nuclear Vowel [a]</th>
<th>Nuclear Vowel [e]</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>F2</td>
</tr>
<tr>
<td>(in Hz)</td>
<td>(in Hz)</td>
</tr>
<tr>
<td>Tone 1</td>
<td>1110</td>
</tr>
<tr>
<td>Tone 2</td>
<td>1080</td>
</tr>
<tr>
<td>Tone 3</td>
<td>1110</td>
</tr>
<tr>
<td>Tone 4</td>
<td>1110</td>
</tr>
</tbody>
</table>

ANOVA were performed to test the variance among the four tones with respect to all the acoustic parameters in Tables 4 and 5. The result show that the differences among the four tones with respect to all these acoustic properties were not significant [for duration: F=1.58, df=(3, 12), N.S.; for intensity: F=0.72, df=(3, 12), N.S.; for F1 of [a]: F=0.06, df=(3, 4), N.S.; for F2 of [a]: F=0.26, df=(3, 4), N.S.; for F3 of [a]: F=3.28, df=(3, 4), N.S.; for F1 of [e]: F=1.33, df=(3, 4), N.S.; for F2 of [e]: F=0.3, df=(3, 4), N.S.; for F3 of [e]: F=2.0, df=(3, 4), N.S.].

IV. DISCUSSION

Whispering, as a naturally occurring mode of communication, is generally highly intelligible. The intelligibility of whispered speech is not hard to understand since except for
the lack of vocal cord vibration, a general drop in intensity, and a general raise of formant frequencies in the vocalic segments, most, if not all, other important acoustic cues for vowels and consonant are kept in the whispered speech signals. In fact, whispering is usually a good illustration of the importance of resonance in speech sound perception—speech is perfectly intelligible even without vocal quality as long as the resonance is intact. A comparison of Figure 1 and Figure 2 below will make this point clear. In Figure 1, we have a spectrogram of one of our contrastive syllables in the test items 飛 [fei] uttered in whisper, while in Figure 2, the spectrogram of a normally phonated version of the same words is displayed. We can easily see that except for the general decrease of intensity (lighter shade of greyness of the spectrogram in Figure 1) and the lack of vocal cord vibration (no regularly spaced vertical striations along the base line of the spectrogram in Figure 1), the general resonance pattern of the whispered word 飛 is highly similar to that of the phonated version in Figure 2. The pattern of the fricative noise of the consonant [f] is almost the same in both spectrograms and so are the patterns of the formants of the vocalic portion [ei] (with a general raise in formant frequencies in the whispered one, though). One can even see quite clearly the rising formant transition in the initial part of [e] (which is due to the influence of the labial consonant [f]) and the smooth transition from [e] to [i] in both spectrograms. Even without any larger contact, a listener will have no difficulty to perceive the sound combination in Figure 1 as [fei]. However, because of the lack of vocal cord vibration, which is the essence of word tones, it may not be easy to determine whether or not the word whispered in Figure 1 is actually 飛, namely, [fei] in Tone 1.
Figure 1. Spectrogram of the Test Item 飛 [fei ʔ] as Produced in Whisper
Figure 2. Spectrogram of the Word 飛 [fei] as Produced by Normal Phonation
The findings of this study is congruent with Miller's (1961) and Wise and Chong’s (1957, as reviewed in Jensen 1958) findings on the perception of word tones in Vietnamese and in Chinese. Both these earlier studies and the present one show that without proper and larger context (e.g., in minimal contrast context or in isolation), word tones are not readily recognizable. In fact, most of the subjects performed only slightly over chance level in their judgement task in these studies (22-58% in Miller's, 62% in Wise and Chong's, and 54% in the present study). It is also interesting to observe that the very speaker who did the recording for our whispered test items could correctly identify only 58% of the test words when the tape was played back to him two weeks after he did the recording--again, only slightly above chance level. Our findings, together with those of Miller (1961) and Wise and Chong (1957), really cast strong doubts on Jensen’s findings (73-88% recognition rate for whispered words in isolation ) and support Miller's claim that very little word tone is transmitted in Vietnamese. If fundamental frequency is the essence of word tone, it only stands to reason that the absences of it would greatly hamper the perception of it.

Our findings that there is no gender and age difference in the subjects' perception capability of whispered tones also strengthen the fact that the essence of word tone is fundamental frequency and there does not seem to be compelling reasons to assume that gender and age should make any difference at all in the process of perception of fundamental frequency variation across time.

Besides the above findings which strengthen our general understanding of the essential perceptual cues of word tones in general. We also find that among the four tones in Mandarin whispered speech, the second and the third tones seem to be perceptually more salient than the first and the fourth tones; and this difference is statistically significant. In our search for clues of this perceptual saliency of the second and third tones over the other two tones, we examine the duration, overall intensity, and formant frequencies of the 16 contrastive syllables of the test items; but we cannot find any significant differences in all these
parameters to which we can link the better performance by the subjects in recognizing the second and third tones. Our failure in this regard implies that either other acoustic parameters should be looked into (e.g., Jensen (1958) pointed out that a vowel tends to have added noise when whispered in an intended rising tone--like Tone 2 in Mandarin) or a replication of the present study should be carried out to re-examine all aspects of the experiment. In whatever way we may look at this aspect of our findings, efforts must be made to seek some kind of explanation since the difference in perception salience of Tones 2 and 3 over the other tones is by no means an accidental finding; both the confusion matrix and the average scores of wrong responses in Table 3 and Table 4 respectively reach the P<0.001 level of significance.

V. CONCLUSION

The recognition of word tones in whispered speech may reveal the significance of fundamental frequency variation in speech sound perception. Phoneticians and linguists have long been interested in this aspect of speech perception but relatively few studies have been done in the literature and a lot of issues remain unresolved. Among them are the question of how well whispered word tones can be recognized, the question of whether or not some tones are perceptually more salient than others, and the question of whether or not whispered word tones are better recognized in some languages than in others. The present study addresses the first two of these questions. Our findings suggest that when fundamental frequency variation is absent as in whispered speech, perception of word tones is greatly hampered if proper and larger linguistic contexts are also excluded. The performance of most subjects in their recognition tasks in our experiment was not good, suggesting that not much word tone is transmitted in whispered speech. This is contradictory to Jensen's (1958) finding and provides some supporting evidence to Miller's (1961) finding, thus bringing some additional evidence to the answer and solution of the first question. As for the
second question, our findings show that Tone 2 and Tone 3 are perceptually more salient in whispered speech; but we fail to find any acoustic correlates to account for this aspect of saliency. In the light of these findings, we suggest that further studies be done to explore the possibility of the existence of other kinds of acoustic correlates and replications of the present experiment be carried out in order to seek better explanations for the perceptual saliency of Tone 2 and Tone 3 in Mandarin whispered speech. We are fully aware of Jensen’s (1958, p.189) claim that "experimental phonetics is incapable of bringing out all the physical correlates to language stimuli." However, we are also convinced that experimental phonetics may very well be one of the very few ways through which this unresolved issue of perceptual saliency could be addressed.

APPENDIX

List of Test Items (non-randomized)

Actual Items of Interest

1. 他的名字叫大威。
2. 他的名字叫大維。
3. 他的名字叫大偉。
4. 他的名字叫大衛。
5. 他的綽號叫大飛。
6. 他的綽號叫大肥。
7. 他的綽號叫大匪。
8. 他的綽號叫大肺。
9. 這兩個字唸大媽。
10. 這兩個字唸大麻。
11. 這兩個字唸大馬。
12. 這兩個字唸大罵。
13. 這兩個字唸大巴。
14. 這兩個字唸大拔。
15. 這兩個字唸大把。
16. 這兩個字唸大壩。

Distractors

1. 老朋友都叫他阿基。
2. 老朋友都叫他阿吉。
3. 這兩個字唸亂掰。
4. 這兩個字唸亂拜。
5. 這兩個字唸亂叫。
6. 他的綽號叫大鍋。
7. 他的家人都叫他阿賢。
8. 他的家人都叫他阿顯。

REFERENCES


Kallail, K. J., and Emanuel, F. W. (1985) The identifiability of isolated whispered and


國語耳語語料中聲調之感知研究
謝國平

提要

聲調的主要聲學特性是語音的基頻，而基頻則取決於發音時聲帶之振動次數。在耳語(whisper)時，聲帶不振動，並無基頻存在。因此，在耳語中，聲調的感知相對相難困難，而其研究亦可顯示基頻在語音感知中的重要性。然而在文獻中，對聲調在耳語的感知研究不多，且有爭議之未決問題則不少。本研究重點在於以實驗方式探討這些未決問題中之二：(1)聲調在國語耳語語料中感知的正確度如何；(2)國語的四聲是否有感知顯著性上的差異。(換言之，在耳語中某些聲調是否比其他聲調更易於感知)

實驗結果顯示：(1)在耳語中，如無語境的幫助，聲調的感知相當困難，受試者只能正確辨識54%的測試聲調。這發現與Jensen(1958)的研究結果相反，但與Miller(1961)的實驗結果相符，顯示在以耳語說出的語料中，並無多少聲調的聲學線索存在。(2)在國語的耳語語料中，第二聲(陽平)及第三聲(上聲)要比其他兩聲更容易被感知；惟對這種感知顯著性上的差異，在本研究的測試語料中並未能發現任何足以解釋此種差異的相關聲學參數。基於上述發現，本文除對耳語語料的聲學特性及其聲調的感知加以討論外，更提出進一步研究的可行方向。

關鍵詞：耳語，基頻，感知顯著性