Mandarin Tone Perception with Temporal Envelope and Periodicity Cues From Different Frequency Regions

Meng Yuan¹, Tan Lee¹ and Sigfrid D. Soli²

¹ Department of Electronic Engineering, The Chinese University of Hong Kong
² House Ear Institute, Los Angeles, CA

ABSTRACT

Temporal envelope and periodicity cues (TEPC) are crucial for speech perception of hearing-impaired people who have poor frequency selectivity. This paper investigates the contributions of TEPCs extracted from different frequency regions to lexical tone perception of Mandarin. Tone identification tests were carried out with tone-contrast ing monosyllabic and disyllabic words. Normal-hearing subjects were recruited in the psychoacoustic experiments with acoustic stimuli that simulate the output of a cochlear implant. The results show that tone identification accuracy with sub-band TEPCs is consistently higher for male voice than for female voice. TEPCs from sub-bands above 1 kHz are found to contribute more to tone identification than those from sub-bands below 1 kHz, especially for male voice. Tone recognition performance can be improved by simply removing the low-frequency TEPCs. The same findings were obtained in our previous study on Cantonese tone perception. This suggests that emphasizing high-frequency TEPCs may be an effective strategy to improve speech perception of tonal languages.

Index Terms—Temporal envelope, temporal periodicity, lexical tone, pitch, tone contour

1. INTRODUCTION

Tone perception plays a vital role in speech perception of tonal languages, e.g., Mandarin and Cantonese. In a tonal language, pitch of voice is used to distinguish one word from another [1]. Each syllable carries a specific lexical tone that is characterized by a distinctive pitch pattern. Pitch is the auditory sensation of the height of a sound. Acoustically it is most closely associated with the fundamental frequency (henceforth abbreviated F0) of the sound wave. In human speech, F0 is determined by the vibration rate of the vocal cords. There are four lexical tones in Mandarin Chinese and six tones in Cantonese, characterized by different pitch patterns over the voiced portion of a syllable [2]. Fig. 1 gives the illustrative pitch patterns of the Mandarin and Cantonese tones. In general, tones are determined by both pitch height and pitch contour. In Mandarin, the four tones are distinguished primarily by their contour shapes, while the identification of Cantonese tones relies heavily on relative pitch level.

Temporal variation of sound wave is useful for human speech perception, especially for hearing-impaired people (HI) who have limited frequency selectivity. Temporal information is the time domain property of speech signal, which is divided into three compo-

This research was partially supported by the Earmarked Research Grants (Refs: CUHK 413405 and CUHK 413507) from the Hong Kong Research Grants Council. CA: myuan@ee.cuhk.edu.hk

Fig. 1. Pitch patterns of (a) Mandarin tones [3] and (b) Cantonese tones [4]. The four tones in Mandarin are numbered from 1 to 4 and described as flat, rising, falling-rising, and falling, respectively. The six Cantonese tones are numbered from 1 to 6 and described as high level, high rising, mid level, low falling, low rising, and low level, respectively.
bands in octave scale: 60 – 500 Hz, 500 Hz – 1 kHz, 1 – 2 kHz and 2 – 4 kHz. The TEPCs were extracted by full-wave rectifying and low-pass filtering the band-passed signals. They were used to modulate noise carriers within the respective sub-bands. The speech materials were Cantonese monosyllabic words. The results showed that TEPCs from high-frequency region (>1 kHz) was more important for Cantonese tone identification than those from low-frequency region. In [12], we investigated the effect of enhancing sub-band periodicity cues on Cantonese tone perception with both monosyllabic or disyllabic word stimuli. In this paper, we extend our studies on Cantonese to Mandarin, which on one hand, is another important Chinese dialect and on the other hand, has a different tone system. The same signal processing method and psychophysical test procedures are adopted. The speech materials include monosyllabic and disyllabic words spoken by one male and one female native Mandarin speaker. It is expected that the major trends observed in our Cantonese studies will also be seen for Mandarin. However, Mandarin tones may exhibit different confusion patterns.

2. METHODS

2.1. Subjects

Five male and five female native Mandarin-speaking listeners participated in the study. Their ages are from 24 to 30 years old. All subjects have normal hearing with pure-tone thresholds better than 25 dB HL at octave frequencies from 125 Hz to 4 kHz in both ears.

2.2. Speech Materials

The speech materials consist of two parts: monosyllabic words and disyllabic words. There are 20 sets of monosyllabic words. Each set contains four words that share the same segmental structure but carry different tones. For example, one set of words were: “母” (mother), “布” (linen), “马” (horse), “曲” (curse). These Chinese characters are all pronounced as the syllable “ma”. The selection criteria of the monosyllabic words include: (1) they are commonly used in daily communication; and (2) they are as many monosyllalan-vowel combinations as possible. One male and one female native speaker were asked to utter the 80 monosyllabic words. Each word was recorded with 5 repetitions and the best one of them was selected for our study.

The disyllabic words were recorded from the same speakers. Given the lexical constraint of Chinese, it is difficult to find disyllabic words that are contrasted only on one of the tones. We therefore decided to include only one pair of contrasting tones in each test set, e.g., “双 /mu2 li4/ (slave)” and “努力 /mu3 li4/ (struggle).” The four Mandarin tones give 6 contrasting pairs and the contrasting tone can be on either the first or the second syllable of the disyllabic words. To provide more segmental variation, we designed two different sets of words for each case of syllable-dependent tone contrast. As a result, there were totally 6 × 2 × 2 = 24 test sets. In addition to the two-tone-contrasting test sets, each test set includes another two distracter words that aim at minimizing the learning effect. The distractor words have exactly the same tone contrast but slightly different segmental compositions. In the above example, the distractor words were “努力 /mu2 li4/ (disability)” and “布 /mu3 li4/ (oyster).” Nearly 90% of the Mandarin phonemes were covered in the test words.

Recordings were made in a sound-treated booth with a high-quality microphone and a desktop computer. The speech utterances were equalized to the same intensity and low-pass limited to 4 kHz. The disyllabic words may exhibit different confusion patterns.

2.3. Speech Processing and Stimuli

In this study, a standard noise-excited vocoder was implemented, acoustically simulating the CIS processing strategy used in CI [13]. The block diagram was shown in Fig. 2. The vocoder starts with a four-channel band-pass filtering process [7] [11]. In each frequency sub-band, TEPC is extracted by full-wave rectification and low-pass filtering of the sub-band signal.

In the re-synthesis stage, the TEPC from each sub-band is first used to amplitude-modulate a band-specific noise carrier signal, i.e., multiplying the TEPC by the noise carrier. Then the modulated signal is band-pass filtered again to ensure that the sub-band output is limited within the respective band. The filtered signal is then scaled to have the same intensity level as the original band-passed signal. Finally, the output is produced by adding the equalized signals. The noise carrier signals are obtained by band-pass filtering a speech-spectrum-shaped noise.

The frequency ranges of the four sub-bands were: 60 – 500 Hz, 500 Hz – 1 kHz, 1 – 2 kHz, and 2 – 4 kHz. All band-pass filters were Elliptical filters. The low-pass filter for TEPC extraction is an Elliptical filter with cut-off frequency of 500 Hz.

Three processing conditions were generated with TEPC-modulated noise carriers from low (60 Hz – 1 kHz), high (1 – 4 kHz) and all (60 Hz – 4 kHz) frequency regions, denoted as LOW, HIGH and ALL, respectively. The parameters and processing methods were adopted from our previous study on Cantonese tone perception [11].

2.4. Psychophysical Procedure

The normal-hearing subjects were seated in a sound proof room and listened to acoustic stimuli presented via a Paired E.A.R. Tone 3A Insert Earphone at 65 dBA. Each subject was required to attend two sessions, in which monosyllabic or disyllabic test materials were used separately. The presentation order of the two types of materials was balanced over all subjects. In each session, the stimuli from one of the two speakers were presented one after another whose order was also randomized and balanced. For each speaker, the unprocessed clean speech was presented at the beginning to let the subjects familiarize with the process and materials. Subsequently the three sets of processed stimuli were presented in randomized order without repetition. For the monosyllabic word tests, two tokens were used in each presentation [11] and the subject was asked to identify both of them. The disyllabic word tests were conducted using a 4-alternative, forced-choice (4-AFC) task. No feedback was provided to the subjects during the tests.
3. RESULTS

Fig. 3 shows the percentage correctness of tone identification with monosyllabic and disyllabic words. The results for male and female voices are displayed separately. The tone scores of LOW, HIGH and ALL are compared. The scores for unprocessed clean speech (ORG) are shown as a reference. Three trends were noted. First, the tone scores under the HIGH and ALL conditions were higher than those under the LOW condition. Second, HIGH was much better than ALL for male voice, while little difference was found for female voice. Third, the scores for male voice were generally higher than those for female voice.

![Fig. 3. Percentage correctness of tone identification with (a) monosyllabic words and (b) disyllabic words. Each pane contains the tone scores attained with the three processing conditions, plus the original unprocessed clean speech. The error bars indicate 95% confidence intervals of the mean scores over all subjects. Chance levels are 25% and 50% for monosyllabic and disyllabic tests respectively.](Image)

The results of tone identification with monosyllabic words and disyllabic words were separately analyzed using a two-way repeated-measures analysis of variance (ANOVA) with the factors of processing condition and speaker. For monosyllabic words, the analysis revealed significant main effects of both processing condition [F(3,27) = 114.12, p < 0.0005] and speaker [F(1,9) = 59.36, p < 0.0005]. A significant main effect on the two-way interaction between processing condition and speaker was also observed [F(3,27) = 27.94, p < 0.0005]. Fisher LSD post-hoc tests showed that, for both female and male voices, the tone scores of HIGH and ALL were significantly higher than those of LOW (p < 0.05). For male voice, the post-hoc comparison showed that HIGH was significantly better than ALL. For all processing conditions, the tone identification scores for male voice were higher than those for female voice. Significant differences between male and female voices were observed for HIGH and ALL.

For tone identification with disyllabic words, ANOVAs revealed significant main effects of both processing condition [F(3,27) = 61.71, p < 0.0005] and speaker [F(1,9) = 13.69, p < 0.005]. The two-way interaction between processing condition and speaker also showed a significant main effect [F(3,27) = 4.44, p < 0.05]. LSD post-hoc tests showed that, HIGH and ALL had significantly better performances than LOW (p < 0.05) for both male and female voices. For male voice, HIGH was significantly higher than ALL. For all processing conditions, the scores for male voice were higher than those for female voice. However, the difference was significant only for the HIGH condition.

The average percentage tone scores over all subjects for different tones were compared in Fig. 4. It was shown that Tone 1 and Tone 2 have higher identification accuracy than Tone 3 and Tone 4 under all processing conditions for both male and female voices. For the HIGH condition with male voice, there was no noticeable difference between the four tones.

![Fig. 4. Percentage correct scores of the four tones with monosyllabic words under different processing conditions.](Image)

4. DISCUSSION AND CONCLUSION

The limited frequency resolution of available cochlear implant systems implies that lexical tone perception depends heavily on the weak temporal envelope cues. Noise-excited vocoder was adopted to simulate the signal processing in cochlear implant systems [7]. In our previous work, Cantonese tone identification was investigated [11]. Sub-band TEPCs below 500 Hz were extracted and used to modulate band-limited noise carriers. Acoustic stimuli were the summation of the modulated sub-band noise carriers. Experimental results showed that TEPCs extracted from the sub-bands above 1 kHz were more important for Cantonese tone perception than those below 1 kHz. This suggests that enhancing TEPCs from the high-frequency region would help tone perception in CIs. The present study is an extension of our Cantonese studies using the same approach. We expected that the importance of high-frequency TEPCs is not specific to Cantonese, but generalizable to other tonal languages, e.g., Mandarin. Due to the different tone systems of Mandarin and Cantonese, we also expected that the contributions of sub-band TEPCs exhibit different patterns in the two dialects.
Our experimental results showed that Mandarin tone identification performance with TEPCs from high-frequency region was significantly better than with TEPCs from low-frequency region. This is consistent with our findings on Cantonese [11]. The reasons may be analyzed in two aspects. First, the frequency regions of the lower frequency noise carriers (60 – 500 Hz and 500 – 1,000 Hz) were overlapped or very close to the frequency range of TEPC (< 500 Hz) such that the noises may mask the useful periodicity information in the TEPCs for tone perception. Second, the higher tone recognition scores may come from the higher periodic fluctuation sensitivities. For normal-hearing people, the pitch perception mechanism is different in different frequency regions whereas the pitch is detected either from the resolved low harmonics in the low frequency region or from the high frequency region where the harmonics are unresolved. In the high frequency region, the pitch is perceived mainly from the temporal periodic fluctuation which represents the F0 of speech. Therefore, the subjects may not effectively utilize the periodic cues in TEPCs from the low frequency region as they do in the high frequency region.

Tone scores attained in the Mandarin tests were higher than those for Cantonese under all processing conditions. This indicates that Mandarin tone perception is a relatively easier task than Cantonese. Indeed, Cantonese tone system is considered more complex, in terms of both the large number of tones and the similarity between the tone contours. The four tones of Mandarin all have very distinctive pitch contours. Our experimental results also showed that, like in Cantonese, Mandarin tone identification performance was consistently much better for male voice than for female voice. This indicates a general trend that the relatively high modulation frequency in female speech might not be represented as well as the low F0 in male speech.

Table 1 and Table 2 showed the tone confusion matrices for HIGH condition in Mandarin and Cantonese, respectively. For the ease of comparison, the numbers shown are all in percentage. For Mandarin, different accuracies were observed among the four tones with Tone 1 being the worst. Tone 3 and 4 showed the highest correctness of tone identification. Fu and Zeng investigated the contributions of temporal cues (amplitude, duration and periodicity) to Mandarin tone recognition [14]. Their results also showed that the accuracy for Tone 1 was the lowest among the four tones. They found that the durations of Tone 3 and 4 were the longest and shortest and thus the duration cues might be helpful to distinguish Tone 3 and 4 from others. In addition, the dynamic ranges of F0 vary a lot among different tones: Tone 4 has the largest F0 range and Tone 1 has the smallest range. The confusion matrix also showed that Tone 1 of Mandarin was easily confused with Tone 2 and Tone 4. For Cantonese, the confusion patterns can be categorized into two groups, which are related to the tone level and tone contour, respectively. There were confusions between tones with close pitch levels, e.g., Tone 1 and 3, Tone 3 and 4, Tone 3 and 6, and Tone 4 and 6, and between the two rising tones, i.e., Tone 2 and 5. The best recognized tone is the high-level tone, which seems to contradict with the observation on Mandarin. This may indicate that relative F0 difference plays a critical role in tone identification.

In summary, this study shows that (i) TEPCs from high-frequency region (> 1 kHz) are more important for Mandarin tone perception than those from low-frequency region; (ii) tone identification performance for male voice was better than for female voice. Delivering and enhancing TEPCs in the high-frequency sub-bands are expected to help improving tone perception and hence speech perception of tonal languages for CI users.

Table 1. Confusion matrix of Mandarin tone identification. The stimulus tones are presented in the y-axis. The response tones are presented in the x-axis.

<table>
<thead>
<tr>
<th>Tone</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>67</td>
<td>14</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>75</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>10</td>
<td>82</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>4</td>
<td>4</td>
<td>83</td>
</tr>
</tbody>
</table>

Table 2. Confusion matrix of Cantonese tone identification. The stimulus tones are presented in the y-axis. The response tones are presented in the x-axis.

<table>
<thead>
<tr>
<th>Tone</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>85</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>78</td>
<td>5</td>
<td>4</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>4</td>
<td>75</td>
<td>4</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
<td>78</td>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>19</td>
<td>7</td>
<td>5</td>
<td>63</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>5</td>
<td>14</td>
<td>8</td>
<td>8</td>
<td>62</td>
</tr>
</tbody>
</table>

5. REFERENCES