MANDARIN TONE IDENTIFICATION WITH F0-FLATTENING PROCESSED SINGLE-VOWELS

Fei Chen

Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China
fchen@sustech.edu.cn

ABSTRACT

Fundamental frequency (F0) contour carries important information for lexical tone identification in a tonal language like Mandarin Chinese. To examine the perceptual contribution of F0 contour, F0-flattening processing has been commonly used in many studies. However, listeners may still capture residual F0 information contained in the F0-flatted stimuli for lexical tone identification. The present work assessed Mandarin tone identification with the F0-flattening processed single-vowels. The vowel stimuli were processed by the STRAIGHT algorithm to replace the original F0 contour with a flattened contour at the mean value of the F0 dynamic contour extracted. Listening experiments showed that under the F0-flattening processed condition, normal-hearing listeners still achieved a more than 50.0% accuracy rate to identify Mandarin tones, and tone-2 was more misidentified as tone-1 than tone-3 and tone-4 were.

Keywords: Mandarin tone identification, fundamental frequency, F0-flattening processing.

1. INTRODUCTION

There are four lexical tones in Mandarin Chinese, namely, the flat, the rising, the falling-rising, and the falling tone (or tone-1, tone-2, tone-3, and tone-4, respectively), each characterized by its pattern in fundamental frequency (F0) variation (or F0 contour) during voiced segments of speech [1]. For a tonal language like Mandarin Chinese, F0 contour carries important information for lexical tone identification and speech perception [2-3]. Many studies have been carried out to understand the perceptual role of F0 contour under various listening conditions, to assess factors affecting Mandarin tone identification, and to design novel speech processing approaches to effectively deliver F0 contour information [e.g., 4-6]. For instance, Chen and Wong recently assessed the segmental contribution to Mandarin tone identification [7]. Liu and Samuel found that Mandarin speakers could flexibly shift to secondary cues such as amplitude in identifying tones [8].

Flattening F0 contour has been used in many studies to assess the effect of F0 contour on Mandarin sentence perception [e.g., 5-6, 9-10]. Several open-source codes are available to implement this F0-flattening processing [11-12]. The underlying principle of these F0-flattening algorithms is to compute the F0 trajectory in voiced segments, and use the averaged F0 value to replace the F0 dynamic contour. However, it is known that much pitch related information impacts lexical tone identification [e.g., 13-15]. Hence, flattening F0 contour may not fully remove the F0 contour information. In other words, the F0 contour information may be preserved in other residual acoustic cues (e.g., amplitude fluctuation), and listeners may still use the residual information (carrying F0 contour cue) for their lexical tone identification and speech perception. Wang et al. used the F0-flattening processed sentences to evaluate the role of F0 contour on Mandarin sentence understanding, and they found that in quiet condition, normal-hearing (NH) listeners still had an almost perfect understanding of Mandarin sentences [5]. Chen et al. used the F0-flattening processing to examine the intelligibility of vowel sentences, which preserved vowel segments and replaced consonant segments with noise/silence [6]. They suggested that, compared with manipulating other cues (e.g., harmonic structure), flattening F0 contour had a minimal effect on the intelligibility of vowel sentences.

All the above-mentioned work was studied at sentence level. At sentence-level speech perception, NH listeners may use additional cues, like language experience and contextual information, to understand processed speech, irrespective of the availability of F0 contour information [e.g., 4]. Due to this interaction with contextual cue, little is known on how and to which extent F0-flattening processing affects Mandarin tone identification. Hence, it is necessary to investigate the effect of F0-flattening processing on Mandarin tone identification at segmental (e.g., vowel) level, which is the purpose of this work. The hypothesis behind this work is that the present F0-flattening processing could only remove the F0-contour information in F0 variation, but could not eliminate other F0-contour

information. Hence, due to the residual F0-contour information preserved, listeners may still identify some tones from F0-flattening processed vowels, rather than identifying all F0-flattening processed vowels as tone-1 (or flat tone).

2. METHODS

2.1. Subjects

Seventeen (ten male and seven female) NH native Mandarin-Chinese listeners participated in the experiment. The subjects' age ranged from 23 to 35 years, and the majority of subjects were undergraduate students at Southern University of Science and Technology. All subjects were paid for their participation.

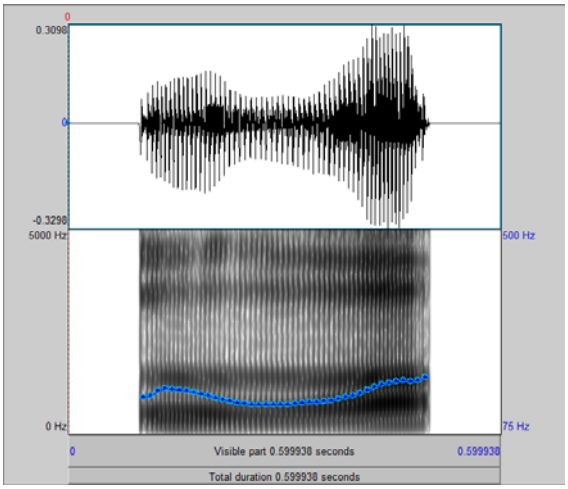
2.2. Materials

One adult male and one adult female native Mandarin-Chinese speakers produced the following six single-vowel syllables in each of the four Mandarin tones (/a/, /o/, /e/, /i/, /u/, /ü/) in a sound-treated booth, resulting in a total of 48 vowel tokens (= 2 speakers × 6 vowels × 4 tones) for Mandarin tone identification. The vowels were recorded at a sampling rate of 16 kHz, and their waveforms were then adjusted to have the same root-mean-square value [16]. The duration of the vowel tokens was normalized [14]. Note that pre-test showed that all participants in this study could correctly identify the tones of all the above single-vowels.

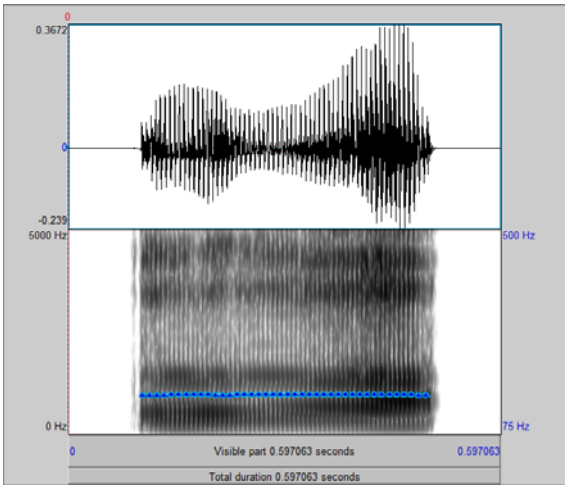
To synthesize the F0-flattening processed stimuli, the F0 dynamic contour of each vowel material was extracted and subsequently replaced by a flattened F0 contour at the mean value of the F0 dynamic contour extracted. More specifically, F0 was extracted every 1 ms in a search range from 40 to 800 Hz. The extracted F0 information was used to control the spectral envelope extraction procedure. The extracted F0 information and the spectral envelope were subsequently used for speech synthesis. In this work, the F0 trajectory was replaced by the flattened F0 contour at the mean value of the F0 dynamic contour extracted (see more in [17]). Through this process, the natural F0 contour was flattened, while other acoustic cues such as formant frequency variations and harmonics were preserved. The Matlab code to implement the above F0-flattening processing (i.e., the STRAIGHT algorithm) is available at [11].

Figure 1 gives the examples of F0-flattening processing. Panel (a) in Fig. 1 shows the temporal waveform and spectrogram of vowel /e/ in tone-3 (by a female speaker), and panel (b) plots the waveform and spectrogram of the F0-flattening processed vowel /e/ originally in tone-3. The F0 contours for the original and F0-flattening processed vowels are shown on the spectrograms. It is seen in panel (a) that F0 contour shows the falling-rising trajectory of tone-3 in Mandarin, while a flat F0-contour is observed in panel (b) due to the F0-flattening processing.

Figure 1: (a) The waveform and spectrogram of a tone-3 vowel /e/, and (b) the waveform and spectrum of the F0-flattening processed tone-3 vowel /e/. The F0 contours are displayed on the spectrograms in the two panels.



(a)



(b)

2.3. Procedure

The experiment was performed in a sound-proof booth and stimuli were played to listeners through a Sennheiser HD 250 Linear II circumaural headphone at a comfortable listening level. Prior to the test, a pre-test showed that all participants could correctly identify the tones of all the F0-unprocessed single-

vowels. In addition, each subject also participated in a 5-min training session and listened to the F0-flattening processed vowels to become familiar with the F0-flattening processed stimuli and the testing procedure. Each subject participated in two testing conditions (i.e., a female voice and a male voice) of Mandarin tone identification, and the order of the two testing conditions was randomized across subjects. Each condition consisted of three presentations of each vowel stimulus spoken by a female or a male talker, and each subject listened to 72 randomized vowel stimuli (= 3 repetitions × 6 vowels × 4 tones) per condition. The Mandarin tone responses were collected with custom software using a computer display of response alternatives and a mouse as a response key. The subjects were allowed to use a repeat key as many times as they wished to repeat the test stimuli during the test. The percentage correct rate for each condition was calculated by dividing the number of tones correctly identified by the total number of vowel stimuli in the testing condition.

Figure 2: Mandarin tone identification scores under all conditions. The error bars denote ± 1 standard errors of the mean. The dotted line indicates the chance level rate (25.0%) for Mandarin tone identification. 's' and 'ns' denote that the difference between the paired scores is significant and non-significantly, respectively.

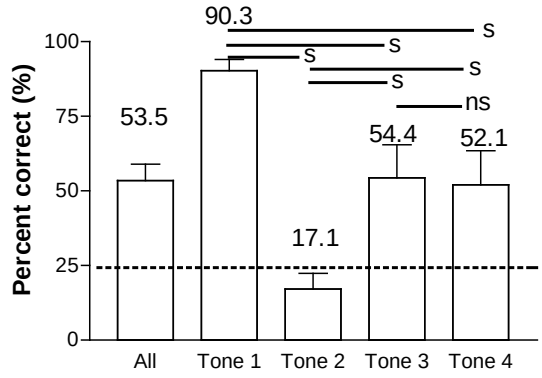


Table 1: Confusion matrix of identifying lexical Mandarin tones.

		Predicted tone			
		Tone 1	Tone 2	Tone 3	Tone 4
Actual tone	Tone 1	90.3%	6.0%	2.1%	1.6%
	Tone 2	75.0%	17.1%	6.7%	1.2%
	Tone 3	24.8%	19.4%	54.4%	1.4%
	Tone 4	26.6%	14.4%	6.9%	52.1%

3. RESULTS

Figure 2 shows the mean (across all subjects, and averaged over the two voice conditions) correct rates of Mandarin tone identification. The mean correct rate 53.5% is much higher than the chance level rate 25.0%, suggesting that NH listeners may use some acoustic cue(s) for assisting their lexical tone identification, although the original F0-contour was replaced with a flattened F0-contour. Figure 2 also splits the tone identification score according to the four tones in Mandarin. First, it is seen that listeners can achieve a high identification rate for tone-1, i.e., 90.3%. On the contrary, for results of tone-2, tone-3 and tone-4, the identification rates are much lower, i.e., 17.1%, 54.4% and 52.1%, respectively. The identification rate for tone-2 is lower than the chance level rate (i.e., 17.1% vs. 25.0%), while the rates for tone-3 and tone-4 are above 50.0% (i.e., 54.4% and 52.1%). Hence, taking the results of four tones together, listeners had correct rate 53.5% for identifying Mandarin tones with F0-flattened vowels. Multiple paired comparisons with Bonferroni correction were run between the identification scores across the four tones in Fig. 2. The Bonferroni-corrected statistical significance level was set at $p<0.008$ ($\alpha=0.05$). Analysis revealed that the score of tone 1 was significantly ($p<0.008$) larger than that of tone 2, tone 3 or tone 4 (noted as 's' in Fig. 2), the score of tone 2 was significantly ($p<0.008$) smaller than that of tone 3 or tone 4 (noted as 's' in Fig. 2), and the score of tone 3 was non-significantly different with that of tone 4 (noted as 'ns' in Fig. 2).

Table 1 shows the confusion matrix of identifying Mandarin tones, which visualizes the effect of F0-flattening processing on tone identification. Each column of the matrix represents the instances in a predicted tone, while each row represents the instances in an actual tone. The correct rates for all tones are located in the diagonal of the matrix, and other cells in the matrix give the error rates of tone misidentification, i.e., misidentifying the actual tone as the predicted tone in the confusion matrix. It is seen that when F0 contour is flattened, tone-2, tone-3 and tone-4 are misidentified as tone-1 with error rates 75.0%, 24.8% and 26.6%, respectively. This shows that when compared with vowels originally in tone-3 and tone-4, those vowels originally in tone-2 are more misidentified as tone-1.

4. DISCUSSION AND CONCLUSIONS

The present work hypothesized that even when F0 contour was flattened with existing F0-flattening processing algorithm (i.e., STRAIGHT), the F0-flattened single-vowels may still carry residual F0 contour information for identifying Mandarin tones. Waveform observation confirmed this hypothesis

(see Fig. 1). The temporal envelope of the F0-flattening processed vowel still largely preserves the original envelope fluctuation without F0-flattening processing. Listening experiments of Mandarin tone identification further verified this hypothesis. Experiment results showed that, instead of identifying all F0-flattening processed vowels as tone-1 or a chance level correct rate of 25.0%, listeners received an overall mean accuracy rate 53.5%. This clearly showed that listeners could still correctly identify some tones with F0-flattening processed vowels. Analysis also showed that, in the context of F0-flattening processing, vowels originally in tone-2 were more misrecognized as tone-1 than those originally in tone-3 or tone-4. This is consistent with early findings. Several work showed that tone-2 was more difficult to identify than tone-3 and tone-4 [e.g., 18-19].

Because this work showed that the F0-contour information could not be fully removed by existing F0-flattening algorithms (e.g., STRAIGHT), caution needs to be taken when interpreting the results of Mandarin sentence recognition in studies using F0-flattening processing [e.g., 5-6]. Those studies aimed to study the auditory processing ability on individual acoustic cue; however, some cues could not be fully separated, e.g., the F0 contour cue. For instance, when using F0-flattening processing to flatten the F0 contour of Mandarin sentences, envelope waveform still carries some residual F0-contour information. Hence, the high recognition of F0-flattening processed sentences could not be interpreted in the absence of F0 contour information. While F0 contour may have minimal effect on sentence perception, which has been reported in several studies, its exact influence needs to be carefully studied. One way to separate the perceptual role of F0 contour is to use tone-1-pronounced speech. Chen et al. used a text-to-speech engine to synthesize sentences with all words pronounced in tone-1 [4]. As all words were pronounced with tone-1 or flat tone, their amplitude fluctuation only carried flat modulation by tone-1. Hence, amplitude fluctuation did not carry the F0-contour information of tone-2, tone-3 and tone-4; and F0 contour is flattened at both frequency variation and amplitude fluctuation. This type of stimuli could be more suitable in studies assessing the perceptual role of F0 contour.

The present work has the following limitations. First, this work assessed the effect of F0-flattening processing on Mandarin tone identification with vowels. Caution needs to be taken when making inference from findings in this study with vowels onto sentences or connected speech processed by F0-flattening. Depending on the context, sentences might show different amplitude fluctuation which is not wholly dependent on tones. In addition, sentences are subject to various phonological changes such as tone sandhi. Second, this study used six single vowels for tone identification. Some single vowels can represent real words in certain tones. For listeners who participated in the experiment, there is a possibility that they mapped what they heard onto some real words to judge the tone of the words, and then participants might not be solely relying on residual F0 information to make their judgments. In other words, some vowels in certain tones are more accurately perceived in part because of the existence of real words in those tones. Third, it is unclear how other factors (e.g., the gender of speaker, initial pitch, voice quality) affect the findings in this work, which warrants further investigation.

In conclusion, the present work assessed Mandarin tone identification with F0-flattening processed single-vowels. Listening experiments showed that when F0 contour was flattened by the existing F0-flattening processing algorithm (i.e., STRAIGHT), listeners can still correctly identify some tone-3 and tone-4. This is partially because other acoustic cues (e.g., amplitude fluctuation) co-vary with F0 contour; and the present algorithms only flatten F0 contour in F0 variation, but not fully remove F0-contour cue. The residual pitch information (e.g., in amplitude fluctuation) may help lexical tone identification with the F0-flattening processed single-vowels.

5. ACKNOWLEDGEMENTS

This work was supported by the Shenzhen High-level Overseas Talent Program (Grant No. KQJSCX2018031911445398), the National Natural Science Foundation of China (Grant No. 61571213), and the Research Foundation of Department of Science and Technology of Guangdong Province (Grant No. 2018A050501001).

6. REFERENCES

[1] Howie, J. M. 1976. Acoustical studies of Mandarin vowels and tones, Cambridge University Press, Cambridge.

[2] Chen, F., Loizou, P. C. 2011. Predicting the intelligibility of vocoded and wideband Mandarin Chinese. J. Acoust. Soc. Am. 129, 3281–3290.

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

[4] Chen, F., Wong, L. L. N., Hu, Y. 2014. Effects of lexical tone contour on Mandarin sentence

intelligibility. J. Speech Lang. Hear. Res. 57, 338–345.
[5] Wang, J. J., Shu, H., Zhang, L. J., Liu, Z. X., Zhang, Y. 2013. The roles of fundamental frequency contours and sentence context in Mandarin Chinese speech intelligibility. J. Acoust. Soc. Am. 134, EL91–EL97.
[6] Chen, F., Wong, S. W. K., Wong, L. L. N. 2014. Effect of spectral degradation to the intelligibility of vowel sentences, in Proc. of the 15th Annual Conference of the International Speech Communication Association (InterSpeech), Singapore, pp. 2002–2005.
[7] Chen, F., Wong, Y. W. E. 2018. Mandarin tone identification with subsegmental cues in single vowels and isolated words. Speech, Language and Hearing, 21, 183–189.
[8] Liu, S. Y., Samuel, A. G. 2004. Perception of Mandarin lexical tones when F0 information is neutralized. Language and Speech, 47, 109–138.
[9] Fogerty, D., Humes, L.E. 2012. The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences. J. Acoust. Soc. Am. 131, 1490–1501.
[10] Laures, J.S., Weismer, G. 1999. Effects of a flattened fundamental frequency on intelligibility at the sentence level. J. Speech Lang. Hear. Res. 42, 1148–1156.
[11] http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/index_e.html
[12] http://www.linguistics.ucla.edu/faciliti/facilities/acoustic/praat.html#noisespeech
[13] Luo, X., Fu, Q. J. 2004. Enhancing Chinese tone recognition by manipulating amplitude envelope: Implications for cochlear implants. J. Acoust. Soc. Am. 116, 3659–3667.
[14] Fu, Q. J. Zeng, F. G. 2000. Identification of temporal envelope cues in Chinese tone recognition. Asia Pac. J. Speech, Language Hearing, 5, 45–57.
[15] Whalen, D.H., Xu, Y. 1992. Information for Mandarin tones in the amplitude contour and in brief segments. Phonetica, 49, 25–47.
[16] Zhou, N., Xu, L. 2008. Lexical tone recognition with spectrally mismatched envelopes. Hear. Res., 246, 36–43.
[17] Kawahara, H., Masuda-Katsuse, I., Cheveigné, A. D. 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech Communication, 27, 187–207.
[18] Zhu, S. F., Wong, L. L. N., and Chen, F. 2014. Development and validation of a new Mandarin tone identification test. Int. J. Pediatric Otorhinolaryngology, 78, 2174–2182.
[19] Krenmayr, A., Qi, B. E., Liu, H. H., Chen, X. Q., Han, D. M., Schatzer, R., et al. 2011. Development of a Mandarin tone identification test: sensitivity index d' as a performance measure for individual tones. Int. J. Audiol., 50, 155–163.