F_0 PERTURBATIONS BY CONSONANTS AND THEIR IMPLICATIONS ON TONE RECOGNITION

Ching X. Xu & Yi Xu

Northwestern University & University of Chicago
xxq@nwu.edu & xuyi@uchicago.edu

ABSTRACT

In this paper, we present a study on variations of fundamental frequency (F_0) in speech production of Mandarin Chinese, and discuss the implications of our results for automatic tone recognition. Target syllables with different consonants and tones were analyzed in continuous speech samples. Compared to the control syllable /ma/ which shows a smooth transition from the onset F_0 value to the current tonal target, syllables with voiceless initial consonants have local raisings of F_0 at voice onset and shorter F_0 trajectories. The effect known as consonant perturbation on F_0 is local and does not alter the original course of known contextual tonal variations. Rather, it is superimposed on other effects, and thus contributes to the appearance of surface F_0 contours. Understanding and consideration of consonant perturbation on F_0 curves therefore should help improve the performance of current tone recognition systems.

1. INTRODUCTION

Fundamental frequency (F_0) is one of the most important variables in speech. It conveys linguistic information and plays a critical role in people's discourse. It is especially crucial in Mandarin Chinese since the pattern of its variation is the main carrier of tone information.

Tone is also known as lexical tone. It is one of the three phonemic elements in Mandarin and serves to distinguish syllables that are composed of the same segments. The same segmental structure may have totally different meanings if it is said with different tones. During conversation, lack of correct tone information usually results in foreign accents or even misunderstanding [13]. This is true not only for communications between people, but also for interactions between human and computers. Therefore, tone recognition has been an attractive topic in automatic speech processing.

Unfortunately, although numerous well-established techniques are successful in recognizing segments, they typically do not have satisfying performance in identifying tones. As a result, tone recognition is often omitted from current systems of continuous speech processing [3]. In these systems, F_0 variations are ignored and tone information can only be guessed at the stage of post language processing by matching sentence hypotheses to detected segment lattices. Obviously, this full dependence on post processing to decode lexical tones is both time consuming and memory consuming. More importantly, ambiguity remains a big issue if the task is a large-vocabulary one, especially when spontaneous speech is of concern [22]. Thus tone recognition is necessary to improve the performance of speech recognition systems and to decrease the cost of time and memory.

The unsatisfactory performance of automatic tone processing reflects the deficiency in our understanding of F_0 contours. Tone information is mainly conveyed by the variations of F_0 contours. Since each tone is synchronized with a corresponding syllable [16, 20], the F_0 trajectory of the syllable is presumably shaped by the underlying tone information. Nevertheless, other communicative elements such as stress and emotion, as well as some articulatory constraints such as coarticulation and consonant perturbation, may also contribute to the F_0 variations [18].

It is difficult to disassemble various factors on F_0 contours at once. Only well-designed experiments, which focus on one or two factors each time while controlling for others, may be able to tease out separate effects [15, 17]. In particular, the influence of unintended constraints needs to be identified to provide a baseline for studying the effects of communicative factors. The current paper presents our effort to identify one of the unintended constraints — consonant perturbation, and its interaction with other factors. Based on our findings, we will discuss potential ways to improve automatic tone recognition.

Consonant perturbation refers to the phenomenon in which a consonant can affect the F_0 movements of adjacent vowels. Research in speech production and perception has confirmed the existence of this effect. It has been found that voiceless consonants give rise to higher following F_0 than voiced counterparts [6, 7, 9], and that F_0 shifts can influence listeners' judgments of consonant voicing [4, 10, 12]. It has also been noticed that F_0 curves may be viewed as combinations of segmental perturbations added onto smooth underlying intonation contours [10]. However, previous studies mostly concentrated on comparing the effects of voiced versus voiceless stops. The interaction between consonant perturbation and other effects on F_0 is seldom addressed. A recent study found that the tone of a syllable with sonorant onset could be identified before any portion of the vowel was heard, while the tone of a syllable with voiceless onset could be mostly correctly identified after the first 20 ms of the vowel was heard [5]. This suggests that F_0 contours during and right after the initial consonant is useful for tone perception, and potentially for automatic tone recognition as well. But the exact relationship between lexical tone and consonant perturbation needs to be well understood before we can make use of it in tone recognition.

2. EXPERIMENT

2.1. Subjects

Seven female native speakers of Mandarin participated as subjects. They were recruited from the Northwestern University community. All subjects had lived in United States for approximately two years at the time of recording. They had no history of speech and language impairments. Their ages ranged from 22 to 30 years old.

2.2. Stimuli

The stimuli were syllables /ma/, /ta/, /t^{h}a/ and /ʂa/, in four tones. Written in the Mandarin Pinyin system, they are “ma”, “da”, “ta” and “sha”, respectively. /m/ does not interrupt the continuous F$_0$ contours in speech [15, 17] and provides a good reference for exploring F$_0$ perturbations by other consonants, so /ma/ served as the control. /t/, /t^h/ and /ʂ/ were chosen because stops (to which /t/ and /t^h/ belong) and fricatives (to which /ʂ/ belong) are the two main kinds of consonants in Mandarin. Moreover, these three sounds are more frequently used in forming syllables than other consonants [13].

Among the vowels, /a/ is the most frequently used in Mandarin, and its pronunciation is relatively stable across syllables [13]. The syllables it forms with the above four consonants can carry each of the four tones, except that /t^{h}a/ with Rising Tone, i.e. /t^{h}a2/, is not phonetically legal in the language. According to Ohde [8], /t^{h}a/ and /p^{h}a/ are consistent in terms of F$_0$ patterns, so /p^{h}a2/ was used as a substitute.

The target tonal syllables were combined into phonetically legal disyllabic words. These words were sorted into two lists. In one list the target syllables were in the first position and in the other the target syllables were in the second position. In both lists, the target tonal syllables had all possible tonal contexts (i.e. four tones). Only the sequence Low-Low was excluded since it is not phonetically different from Rising-Low due to tone sandhi [1].

The disyllabic words were incorporated into two carrier sentences. They were "wo3 lai2 shuo1 ____ zhe4 ge4 ci2"[1] ("I say the word ____") and " wo3 lai2 zhao3 ____ zhe4 ge4 ci2" ("I look for the word ____"). Only the syllables preceding the target words were different in the two conditions. The first was a High Tone, while the second a Low Tone.

2.3. Procedure

The recording was made in a soundproof booth, one subject per session. The subject was seated comfortably in front of a computer screen and wore a head-mount microphone which was approximately 2 inches away from the subject's mouth.

At the beginning of the session, a set of instructions was displayed on the screen, directing the subject to read the target sentences aloud at a normal speaking rate and with stress on the target words. After reading the instructions, the subject went through a series of practice trials to become familiar with the target words before the real trials. In each trial, an experimental sentence in Chinese characters appeared on the screen, and the subject read aloud the sentence as instructed. The subject had control of two buttons on the screen: “Next” and “Again”. Thus, she could choose either to continue with the next sentence when she had successfully finished the current one, or to repeat the current sentence if she had made a mistake.

Each experimental sentence was presented to the subject five times and the order of the repetitions was randomized. There were a total of 1110 sentences (111 target words * 2 carrier sentences * 5 repetitions). Each session lasted for about one hour. The subjects' speech was digitized by SoundEdit (Macromedia Inc.), with 22kHz sampling rate and 16-bit accuracy.

2.4. Data Extraction and Measurement

F$_0$ curves of target syllables were extracted using a method that combined automatic vocal cycle detection and manual rectification [15, 17].

Using the ESPS signal processing software package (Entropic Inc.), every vocal cycle in the target words was marked by the program EPOCHS in the package. Then the marks were manually checked and corrected. At the same time, the onset and offset of each segment of the target syllables were manually labeled using the program XLABEL in XWAVES.

The markings of vocal pulses between the segment labels were processed to extract the F$_0$ value of each vocal cycle. The F$_0$ trajectories were smoothed individually by applying a three-point median filter. Finally, the five repetitions of each syllable by each subject were averaged to remove random variations.

3. RESULTS

The effects of consonant perturbation on F$_0$ contours of Mandarin syllables in four tones are illustrated in Figure 1. Each curve represents an average across 5 repetitions, 4 posterior tonal context and 7 subjects. Target syllables were produced in the first position of disyllabic stimulus words and with a high preceding tone in the carrier sentence. All curves are aligned to the syllable offset.

In each tone condition, /ma/ exhibits the longest F$_0$ tracing which shows a smooth transition from the onset value to the current tonal target throughout the whole syllable. In contrast, the syllables with voiceless initial consonants have local raisings of F$_0$ at voice onset and shorter F$_0$ trajectories. Nevertheless, all the F$_0$ curves with the same tone converge to the same pattern towards the syllable offset. If we ignore the local raisings, /ta/, /t^{h}a/ and /ʂa/ have similar F$_0$ curves as /ma/. Since the visible F$_0$ curves of different syllables start at different times relative to the syllable offset, they fall into different phases of the transition from the syllable onset to the current tonal target.

Figure 2 illustrates an example of the relationship between consonant raising and tonal coarticulation. Each curve represents an average across 5 repetitions, 2 carrier sentences and 7 subjects. All target syllables were produced in the second position of the disyllabic words. As in Figure 1, all curves are aligned to the syllable offset.

The onset F$_0$ value of /ma/ apparently inherits the offset F$_0$ value of its preceding tone, as was described by Xu [15]. When the preceding tone is High or Rising, the F$_0$ contour of /ma/ starts at a much higher value than when the preceding tone is Low or Falling. The differences due to the previous tone decrease over time and almost vanish at the syllable offset during the continuous transition from the onset value to the current tonal target. The syllables with voiceless consonants do not have direct F$_0$ inheritance from the preceding tone, but they do show the same pattern of assimilated carryover effects.

[1] The numerals 1, 2, 3 and 4 represent High, Rising, Low and Falling Tone, respectively.

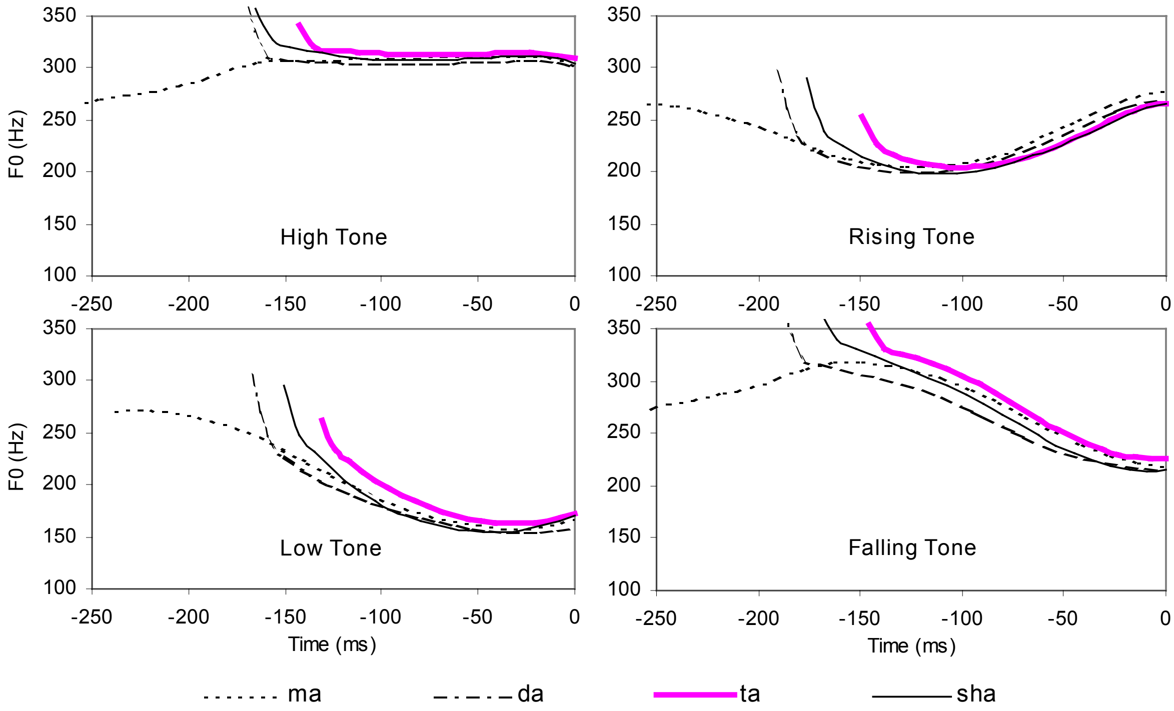

Figure 1: Effects of consonant perturbation on the F_0 contours of Mandarin syllables in four tones

Furthermore, each voiceless consonant superimposes a local raising onto the contextual tonal pattern.

4. DISCUSSIONS AND CONDLUSION

The F_0 perturbations by different voiceless consonants may originate from different sources of constraints. Due to space limitation in the present paper, that issue will be discussed elsewhere. Here we focus only on results that are relevant to tone recognition.

Although consonant perturbation and tonal coarticulation both contribute to the shape of F_0 contours, they mainly affect the beginning portion of a syllable. These effects do not prevent various F_0 curves with the same tone from converging to the same pattern towards the syllable offset. The robust F_0 pattern in each tone condition is consistent with the pitch target approximation model we proposed previously [14, 20]. According to our model, each tone is associated with an underlying pitch target and the pitch target is implemented continuously and asymptotically throughout the duration of the corresponding syllable. As a result, the ending portion of an F_0 curve is expected to be more tone-pertinent than the beginning portion. Therefore, we can use the most tone-pertinent portions of F_0 curves, rather than other portions or the whole curves, for automatic tone recognition. There have been efforts in this line, which show encouraging results [11, 21].

Interestingly, it has recently been demonstrated that listeners could mostly identify a tone as soon as the first 20 ms of the vowel was heard even with the presence of voiceless consonant [5]. For one thing, this finding adds further weight to the claim that listeners make full use of the signal given to them [12]. For the other, it demonstrates that the starting portion of an F_0 curve does contain tone information. Thus automatic tone recognition may also fair better by taking advantage of this information rather than ignoring it. By including the less tone-pertinent portions of F_0 curves into speech processing, we may potentially get more information about the consonants to be recognized as well.

Besides the F_0 curve, the durations of a syllable and its segments also carry information related to tone, since the effects on F_0 and on its time course vary closely with each other. In both Figure 1 and 2 we can see the length differences of F_0 contours across conditions. These differences agree with a previous study on the durations of segments and syllables [2]. It seems that during the interval of voiceless consonants, there are also F_0–related adjustments going on, so that visible F_0 curves that start at different times fall into different phases of the transition from the syllable onset F_0 to the current tonal target. This suggests that time is an important factor in shaping the surface F_0 contours. Time course may become more crucial when syllables get shorter at fast speech rate and the tonal targets cannot be fully implemented because the speaker cannot exceed the maximum speed of pitch change [19]. Therefore, it may be beneficial to give time course serious consideration in automatic tone recognition.

To conclude, our study of F_0 perturbations by consonants in Mandarin and their interaction with contextual tonal variations reveals that the consonant effect is superimposed on the course of contextual tonal patterns. The voiceless portion of a consonant may hide part of the F_0 contours and introduce brief local F_0 raising, but it does not alter the general course of the F_0 transitions toward the underlying tonal targets. These findings

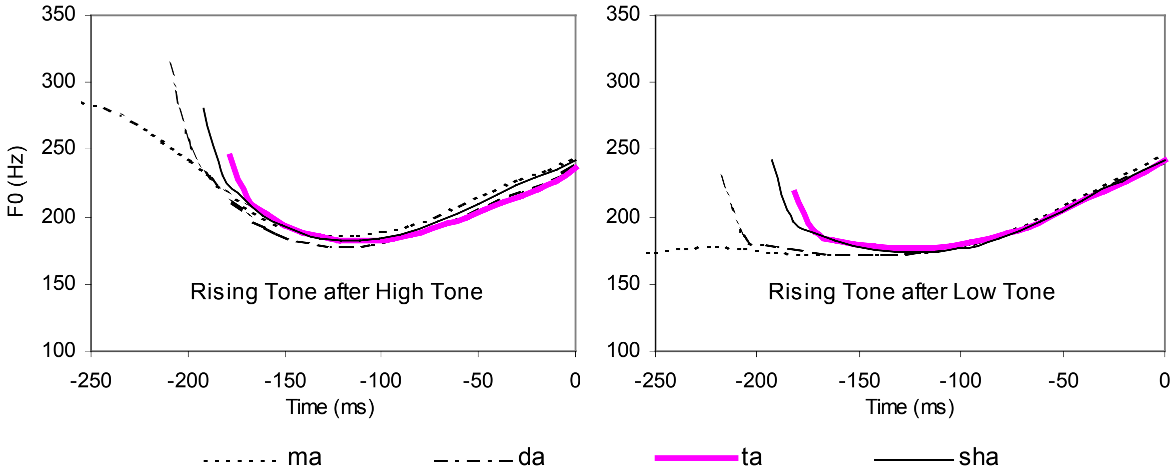

Figure 2: Effects of consonant perturbation on the F_0 contours of Rising Tone after High Tone and Low Tone

suggest that automatic tone recognition may be improved by taking into consideration both the most tone-pertinent portion of the F_0 contours near the end of a syllable and the early F_0 in a syllable which reflects the contributions of the preceding tone, the current tone, and the initial consonant.

5. REFERENCES

[1] Y.R. Chao, *A grammar of spoken Chinese*, Univ. of California Press, Berkeley, CA, 1968.

[2] L. Feng, "Duration of initials, finals and tones in Beijing dialect", *Working Papers in Experimental Phonetics*, Peking Univ. Press, Beijing, 131-195, 1985.

[3] Stephen W. K. Fu, C. H. Lee & Orville L. Clubb, "A survey on Chinese speech recognition", *Communications of COLIPS Proceedings*, Vol. 6, No. 1, P96001, 1996.

[4] J. M. Hombert, "Consonant types, vowel quality, and tone", in V. A. Fromkin (Ed.), *Tone: A linguistic survey*, Academic Press, New York, pp. 77-111, 1978.

[5] C.-Y. Lee, *Lexical Tone in Spoken Word Recognition: A View from Mandarin Chinese*, Ph.D. Dissertation, Brown University, 2001.

[6] I. Lehiste, "Suprasegmental features of speech", in N. J. Lass (Ed.), *Principles of experimental phonetics*, Mosby, Boston, pp. 226-244, 1996.

[7] J. J. Ohala, "The production of tone", in V. A. Fromkin (Ed.), *Tone: A linguistic survey*, Academic Press, New York, pp. 5-39, 1978.

[8] R. Ohde, "Fundamental frequency as an acoustic correlate of stop consonant voicing", *J. Acoust. Soc. Am.* 75, 224-230, 1984.

[9] J. P. H. v. Santen & J. Hirschberg, "Segmental effects on timing and height of pitch contours", *Proceedings of the 3th International Conference on Spoken Language Processing*, pp. 719-722, 1994.

[10] K. Silverman, "F_0 segmental cues depend on intonation: the case of the rise after voiced stops", *Phonetica*, 43, 76-91, 1986.

[11] X. Wang & J. Iso-Sipilä, "Low complexity Mandarin speaker-independent isolated word recognition", *Proceedings of the 7th International Conference on Spoken Language Processing*, pp. 1589-92, 2002.

[12] D. H. Whalen, A. S. Abramson, L. Lisker & M. Mody, "F_0 gives voicing information even with unambiguous voice onset times", *J. Acoust. Soc. Am.* 93, 2152-9, 1993.

[13] Z. Wu, *Essentials of modern Chinese phonetics (in Chinese)*, Foreign language printing house, Beijing, 1992.

[14] C. X. Xu, Y. Xu & L.-S. Luo, "A pitch target approximation model for F0 contours in Mandarin", *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, 2359-2362, 1999.

[15] Y. Xu, "Contextual tonal variations in Mandarin", *Journal of Phonetics*, 25, 61-83, 1997.

[16] Y. Xu, "Consistency of tone-syllable alignment across different syllable structures and speaking rates", *Phonetica*, 55, 179-203, 1998.

[17] Y. Xu, "Effects of tone and focus on the formation and alignment of F_0 contours", *Journal of Phonetics*, 27, 55-105, 1999.

[18] Y. Xu, "Sources of tonal variations in connected speech", *Journal of Chinese Linguistics*, Monograph series #17, 1-31, 2001.

[19] Y. Xu and X. Sun, "Maximum speed of pitch change and how it may relate to speech", *J. Acoust. Soc. Am.* 111, 1399-1413, 2002

[20] Y. Xu & Q. E. Wang, "Pitch targets and their realization: Evidence from Mandarin Chinese", *Speech Communication* 33: 319-337, 2001.

[21] J.-S. Zhang, G. Kawai & K. Hirose, "Subsyllabic tone units for reducing physiological effects in automatic tone recognition for connected Mandarin Chinese", *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, 2367-2370, 1999.

[22] J.-S. Zhang & K. Hirose, "Anchoring hypothesis and its application to tone recognition of Chinese continuous speech", *ICASSP 2000*, Istanbul, Turkey, Vol. 3, pp. 1419-1422, 2000.