SECOND-DEGREE POLYNOMIAL APPROXIMATION OF MANDARIN CHINESE LEXICAL TONE PITCH CONTOURS – A PRELIMINARY EVALUATION

Tomasz Kuczmarski[1], Daniel Duran[2], Norbert Kordek[1], Jagoda Bruni[2]

1. Institute of Linguistics, Adam Mickiewicz University, Poznań
2. Institute for natural language processing, Stuttgart University, Stuttgart

Abstract. The current paper presents a preliminary evaluation of a second-degree polynomial pitch stylization method for MC cited lexical tones. This study was devised to verify methodological assumptions for a subsequent work where a systematic manipulation of the F0 curve in MC syllables will be used to study the perceptual Magnet Effect. For this purpose, a number of MC syllables representing various phonological templates were chosen from a single speaker corpus. Stylized pitch curves were resynthesized and compared with their natural counterparts in a discrimination experiment. The results of native speakers' judgments show that the approximation method is adequate for the desired application.

1 Introduction

The present study attempts to describe a second-degree polynomial F0 approximation method for the Mandarin Chinese (MC) citation tones and the perceptual acceptability of the approximated F0 contours as judged by the native speakers of Mandarin.

Any inquiry into the phonological system of MC usually entails more or less explicit reference to the syllabic system. From the typological perspective it is well known that MC is a tonal language with a relatively small number of 4 tones, a small number of just over 400 tonally undifferentiated syllables, well defined syllable boundaries and constraints on the syllable structure, and finally, with a relatively simple word structure in terms of the number of syllables per word. The common feature of all syllables is an obligatory pitch contour – the tone. Tones are one of the basic phonological strategies of MC for distinguishing the segments on the semantic level. The tones are usually described as: T1 – high-level (55), T2 – high-rising (35), T3 – dip-rising (214), T4 – falling (51), where the numbers in parentheses indicate the change in pitch over time. 4 tones make the ubiquitous homonymy manageable, but ca. 1300 syllables with tones (out of 1600 theoretically possible) in a language in which words are mostly bisyllabic puts the speaker in a situation where the production of tones is of key importance.

We adopt Duanmu's (2007: 82) model of syllable structure, sound representation and tone bearing unit. The syllable (σ) structure is represented by onset (O), Rhyme (R) and timing slots (X):

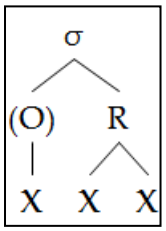

Figure 1. Syllable structure representation after Duanmu (2007)

Different types of syllables may be differentiated in terms of onset, nucleus and coda (Duanmu, 2007: 82):

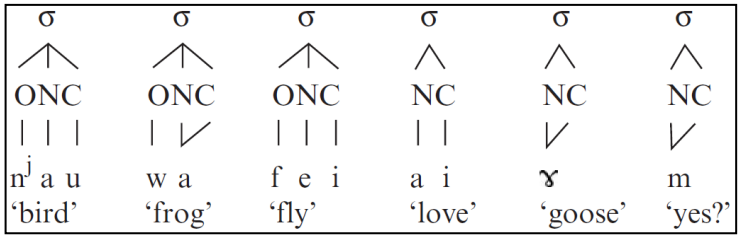

Figure 2. Possible syllable types after Duanmu (2007)

The treatment of consonant + glide (CG) forms as a single sound renders 12 possible syllable structures in terms of phonetic categories: C, V, GV, VC, CV, VG, CVG, GVG, CVC, CGV, CGVC, CGVG.

In Duanmu's model the tones are linked to the rhyme segments and the syllable onset is excluded from tonal functions. The representation of the falling contour tone in this model ('H' and 'L' represent level high and level low tones that constitute the contour falling tone) is presented below (Duanmu, 2007: 234):

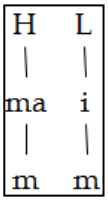

Figure 3. Representation of high and low level tones after Duanmu (2007)

The current study presents a preliminary evaluation of second-degree polynomial functions as a model of Mandarin Chinese (MC) lexical tone F0 contours. The authors' main aim is to verify a set of methodological assumptions for a subsequent work, where systematic manipulation of synthetic F0 contours will be used to study the perceptual Magnet Effect (Lacerda, 1995) in MC. Apart from the theoretical considerations the findings of the study are aimed to facilitate MC teaching and learning process as a part of a planned Computer-Aided Pronunciation Training system.

2 Methodology

1.1 Speech Material

For a preliminary evaluation of the approximation method 60 phonetically diversified syllables representing all four MC tones in 15 syllable templates (Table 1) were selected from a single speaker speech corpus (Shih, 2010).

Syllable template	**Representative syllable *pinyin***
C	*Not available in speech corpus*
V	*a*
GV	*wa*
VC	*an*
CV	*zha*
VG	*ei*
VV	*ao*
CVV	*rao*

CVG	*bai*
CVC	*lan*
GVG	*wei*
GVC	*wen*
CGV	*tuo*
CGVC	*chuang*
CGVG	*kuai*
CGVV	*piao*

Table 1. Syllable templates and examples from the corpus.

Speech data was segmented and annotated by trained phoneticians using a set of 4 labels; consonant (*C*), vowel (*V*), glide (*G*) and an extra label (*?*) representing any kind of glottalization, such as a vocal creak. Fundamental frequency of all utterances was estimated from the acoustic signal using Praat's pitch detection algorithm (Boersma 1993) with a 10ms frame length. The *least squares* method was employed to calculate best fitting second-degree polynomial function parameters for F0 values extracted from vowel (*V*) and glide (*G*) segments only. F0 values occurring within consonant (*C*) and glottalized segments (*?*) were ignored based on the assumption that syllable onset is excluded from tonal functions (Duanmu, 2007). It was also anticipated that this approach would help minimize the influence of incorrectly estimated F0 values at syllable boundaries and glottalized parts of the utterance. The resulting approximation functions were employed to calculate new F0 values at 10ms intervals for all segments except consonants, where the originally extracted values were kept. The resulting contours were resynthesized using Praat's overlap-add method (Moulines &

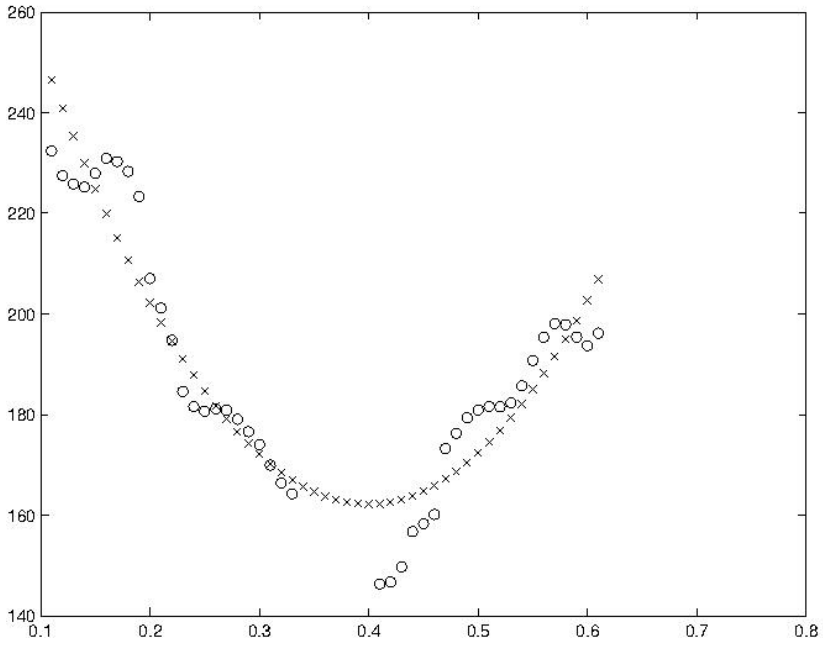

Figure 4. *Rao*3 syllable -- extracted (o) and approximated (x) F0 values

Charpentier, 1990). Figure 4 shows data plot of the approximated and original F0 values for *rao*3 syllable.

1.2 Perception experiment

In order to test the acceptability of the syllables with the resynthesized pitch contours, a discrimination experiment was carried out using Praat's *ExperimentMFC* facility. The 60 original syllables were paired with their resynthesized counterparts in both possible orders. Along with the pairs of identical original and resynthesized syllables, the total number of stimuli summed up to 240. The stimuli were presented to the participant via ordinary headphones in a quiet room in random order using the *PermuteBalancedNoDoublets* randomization method. After hearing a stimulus consisting of two syllables, the participant had to click on a button on the screen indicating whether she heard a difference or not. The experiment was self paced but each stimulus was played only once.

In a first run, two native speakers participated in the study: one female and one male, both students, around 30 years old without any known hearing impairments. The participants were not paid and they did not receive credits.

3 Results

Tables 2 and 3 demonstrate results from the perception experiments obtained for all kinds of syllable pairs (natural–natural, synthetic–synthetic and mixed) obtained for the male (M) and female (F) speaker respectively.

	general (all types)		natural		synthetic		mixed	
(M)	Total	%	Total	%	Total	%	Total	%
correct	133	55%	48	80%	50	83%	35	29%
incorrect	107	45%	12	20%	10	17%	85	71%
Total	240		60		60		120	

Table 2. Correctness of answers for all types of *syllable* pairs - male speaker.

	general (all types)		natural		synthetic		mixed	
(F)	Total	%	Total	%	Total	%	Total	%
correct	122	51%	59	98%	60	100%	3	3%
incorrect	118	49%	1	2%	0	0%	117	98%
Total	240		60		60		120	

Table 3. Correctness of answers for all types of *syllable* pairs - female speaker.

For the male speaker results indicate that recognition of synthetic syllable pairs was slightly better than the natural ones, whereas majority of responses provided for the mixed, natural-synthetic syllable types are incorrect.

The female speaker's responses demonstrate full correctness of answers in case of synthetic syllable pairs and very low rate of correct ones for the mixed pairs.

To sum up, the results indicate a mismatch between the perception of natural stimuli (natural-natural syllable pairs) and fully synthetisized pairs. The recognition seems to be rather balanced for both speakers: for the male participant the correctness for natural syllables is 80% and for synthetic ones 83%. However correct perception of the mixed pairs is only at 29% level. Similarly, the female speaker's responses show relatively comparable tendencies - 98% correct answers for the natural stimuli syllables and 100% for the synthesized ones.

Correctness rate for the mixed syllable types is even lower than in case of the male speaker (3%).

4 Conclusion

The proposed method was developed for the purpose of a subsequent work, where systematic manipulation of synthetic F0 contours will be used to study the perceptual Magnet Effect (Lacerda, 1995) in MC cited lexical tones. Therefore a simple model with few easily-controlled parameters was sought after as a representation of the pitch curve. F0 contours of all MC cited lexical tones resemble a parabola to some extent. That is why second-degree polynomial functions were chosen. However, an experimental evaluation was needed to determine whether this stylization method is not over-simplistic and if it has no perceptual consequences on its own. The results of the current study demonstrate that native MC speakers were virtually unable to differentiate between the natural and stylized pitch contours. On the other hand, it is expected that significant manipulations of the pitch curve, i.e. resynthesizing a T1 syllable with a stylized T3 pitch curve may yield bad results. This might happen due to a number of factors, e.g. the currently used resynthesis method does not provide a tool for a manipulation of the spectral cues for the perception of pitch. Therefore, the listeners might perceive the resythesized syllables as very unnatural, what in turn, might affect the results of the study. This problem is yet to be solved.

Acknowledgments

This research was in part funded by the German Research Foundation (DFG), grant SFB 732, A2 (for authors Bruni and Duran).

References:

Boersma, P. 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, IFA Proceedings, 17:97-110.

Duanmu, S. 2007. The Phonology of Standard Chinese. Oxford University Press.

Duanmu, S. 2008. Syllable structure. The limits of variation. New York: Oxford University Press.

Huang, L. M. 1992. Remarks on the Phonological Structure of Mandarin Chinese. Bulletin of National Taiwan Normal University, 37: 363-383.

Lacerda, F. 1995. The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory. Proceedings of the XIIIth international congress of phonetic sciences, 2:140-147.

Moulines, E., Charpentier, F. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Communication, 9: 453-467.

Shih, Ch. 2010. An Adaptive Training Program for Tone Acquisition. Proceedings of Speech Prosody 2010, Illinois, USA.

Zhang, J. 1996. On the Syllable Structures of Chinese Relating to Speech Recognition. Proceedings of the 4th International Conference on Spoken Language, 2450-2453.