Quantitative Analysis of Tone Coarticulation in Mandarin

Hussein Hussein[1,2], *Hansjörg Mixdorff*[1], *Hue San Do*[1] *and Rüdiger Hoffmann*[2]

[1] Department of Computer Sciences and Media, Beuth University of Applied Sciences, Berlin, Germany
[2] Laboratory of Acoustics and Speech Communication, Dresden University of Technology, Dresden, Germany

{hussein, mixdorff, hsdo}@beuth-hochschule.de
{hussein.hussein, ruediger.hoffmann}@ias.et.tu-dresden.de

Abstract

The current paper examines the effect of tone coarticulation in Mandarin on the amplitude and duration of tone commands of the Fujisaki model and whether declination needs to be taken into account when synthesizing *F0* contours of Mandarin. Based on a corpus of short sentences mean parameters of the Fujisaki-model were calculated for the 15 combinations of Mandarin tones. The resulting smoothed *F0* contours differ from the canonical shapes due to tonal coarticulation. Results of averaged parameters suggest that sequences of tone commands with the same polarity can usually be merged into a single tone command because their tone command amplitudes *At* are very similar. *T2* for the first and *T1* for the second command in these sequences are also very similar, so they can be set to the same value. As a consequence, tonal combinations can be interpreted as sequences of tone switches between high and low tones, considerably simplifying the modeling. It was also found that for most utterances phrase commands of magnitude *Ap* greater 0 occurred, indicating that the phrase component should be taken into account when analyzing and synthesizing of *F0* contour of Mandarin.

Index Terms: Mandarin tone, Fujisaki model, prosodic analysis

1. Introduction

The current study examines the influence of tonal coarticulation in Mandarin on the parameters of the Fujisaki model. It was conducted during the on-going development of a Mandarin training system for German learners within a three-year project funded by the German Federal Ministry of Education and Research [1][2].

It is commonly known that Mandarin is a tone language and hence the tonal contour of a syllable changes its meaning [3]. The most important acoustic correlate of tone is *F0*. Mandarin has four syllabic tones and a neutral tone in unstressed syllables. In citation forms of monosyllabic words the tonal patterns are very distinct, but when several syllables are connected, *F0* contours observed vary considerably due to tonal coarticulation. We observed that the acquisition of tonal patterns of poly-syllabic words is much more difficult than of mono-syllabic words [1].

In the framework of our Mandarin training system one mode of feedback projected is the resynthesis of tonally corrected learner utterances. Although a direct transplantation of native speakers' *F0* contours onto the learner's imitation is a possible option, the time-warping which will have to be applied to the native speaker's contour will also affect the *F0* slopes. Furthermore, only native utterances stored in the system could be employed. Therefore we explore the alternative of modeling the *F0* contour in order to be able to overcome the drawbacks mentioned.

The well-known Fujisaki model is a parsimonious method for parameterizing *F0* contours in speech synthesis for intonation analysis and intonation generation [4]. The model reproduces a given *F0* contour by superimposing three components: a speaker-individual base frequency *Fb*, a phrase component and an accent component in stress-timed languages or a tone component of positive and negative polarities in tone languages. As previous studies showed Mandarin tone can be represented by prototypical *F0* contours [5] and requires tone commands of positive and negative polarity (see Table 1). The phrase component results from responses to impulse-wise phrase commands associated with prosodic breaks. Phrase commands are described by their onset time *T0*, magnitude *Ap* and time constant α (time constant of the phrase control mechanism). The tone component results from step-wise tone commands associated with syllable-tones. In [5] tone commands are described by on- and offset times *T1* and *T2*, amplitude *At*, time constant β_p of the tone control mechanism for positive commands, time constant β_n of the tone control mechanism for negative and constants γ_p (relative ceiling level of positive tone components) and γ_n (relative ceiling level of negative tone components). The use of a common value for both constants β and γ (20/s and 0.9, respectively) irrespective of the polarity of tone command is acceptable [6].

Table 1. *Mandarin tones with prototypical tone command assignment [6].*

Tone	*F0* contour	Tone commands assigned
1	high	positive
2	rising	negative/positive
3	falling-rising	negative
4	falling	positive/negative

With respect to Mandarin, the Fujisaki model has been criticized for requiring too large a number of parameters to account for the *F0* contour associated with a single syllable [7]. This is indeed true if one assumes that each syllable is associated with two independent tone commands, each command with its respective *T1*, *T2*, *At*, as well as β and γ. However, previous experiments suggested that Tone 1 and Tone 3 are typically associated with a single tone command of positive and negative polarity, respectively. Furthermore, depending on the underlying tones, commands of the same polarity have been observed to be stretching across at least two syllables, hence the syllabic tone commands are actually concatenated into longer tonal gestures [8].

The current paper is intended to study in detail the effect of tonal coarticulation on the duration and amplitude of tone commands in syllabic sequences, as well as quantify the

timing differences observed in different tonal combinations of two consecutive syllables. We hypothesize that the description of tones using the Fujisaki model requires fewer parameters than projected in the prototypical tone assignment. Instead, a sequence of tones in Mandarin can be efficiently described as a sequence of characteristic tone switches between high and low *F0* targets with specific amplitudes and timings with respect to the underlying syllabic grid.

Another area we wish to examine is whether or not declination in Mandarin needs to be taken into account in *F0* contour synthesis. The target approximation model presupposes that the observable *F0* contour is entirely derivable from the underlying tones of an utterance [7] and that *F0* down-drift is merely a local phenomenon. In contrast, work by Tseng et al. seems to indicate that the phrasal *F0* contour is an important cue for discourse organization in Mandarin [9]. In the context of the Fujisaki model, we expect the absence of declination to result in phrase command magnitudes close to 0.

2. Experiment Method

2.1. Speech Material

The data used in this experiment consists of recordings from native speakers of Mandarin, all of whom employees at *iFlyTek* company, Hefei, China. The underlying sentences were designed and collected for the purpose of comparison with utterances from German learners of Mandarin.

The data was recorded with a sampling frequency of 16kHz and a resolution of 16 bit. The corpus consists of 62 sentences. The corpus comprises six different sentence types: declarative sentences, yes/no questions, wh-questions, and rhetorical questions, imperative and exclamatory sentences. They contained both monosyllabic and disyllabic words, with a minimum of two and a maximum of 14 syllables. The sentences were provided in Chinese characters on a computer screen and read aloud by the subjects. The sentences were produced by 20 native speakers of Mandarin, ten males and ten females, yielding a total of 1240 utterances. Results on the comparison between German and Chinese subjects were reported in [2].

2.2. Method of Analysis

The *F0* contour reflects the tone on the syllable level. Therefore, the data was forced-aligned on the syllable and phone-levels using the automatic speech recognition (ASR) system in a forced alignment mode. The used ASR system is part of an automated proficiency test of Mandarin [10].

The *F0* contours were calculated using the *Praat* [11] algorithm with a step of 10msec and different standard settings of the minimum and maximum parameters of *F0* for male (100 and 350 Hz) and for female speakers (120 and 450 Hz). It was found by checking the *F0* contours that some speech signals do not have *F0* values in the voiced segments. Therefore, the *F0* contours for such signals were recalculated. The modified parameters of the minimum and maximum *F0* were (50 and 250 Hz) for male and (80 and 400 Hz) for female speakers. The *F0* contours were checked and corrected using the *Praat Pitch-Editor*. The Fujisaki parameters were estimated automatically using the algorithm [8] and if necessary corrected using the interactive tool *FujiParaEditor* [12].

2.3. Extraction of Fujisaki Model Parameters

In the automatic approach to estimate the Fujisaki parameters for Mandarin [8] first the *F0* contour is interpolated and smoothed using a third-order spline. A high-pass filter is then applied to extract the 'high frequency contour' (HFC) from the smoothed *F0* contour. The HFC contains the tone commands. The HFC is subtracted from the smoothed contour, yielding a 'low frequency contour' (LFC) from which the phrase commands are extracted. The parameters of phrase commands were initialized. In order to initialize the number of *T1*, *T2* and polarity of tone command the smoothed *F0* contour was subdivided into segments with positive or negative gradient, respectively. In the derivative segments, the local maxima are searched which represent the inflection points of the smoothed contour. The detected inflection points on *F0* subcontours with rising slope correspondent to the offset of a negative tone command and the subsequent onset of a positive tone command. The offset of a positive tone command and the subsequent onset of a negative tone command are associated with the inflection points on *F0* subcontours with falling slope. *At* is initialized as the positive or negative maximum in the HFC between the initialized values of *T1* and *T2*. The time constants α and β are set to 2/s and 20/s, respectively.

The Analysis-by-Synthesis procedure is performed in three steps to optimize the initial parameters iteratively by applying a hill-climb search. At the first step, phrase and tone components are optimized separately using the target contours LFC and HFC, respectively. Then, phrase, tone component and *Fb* are optimized jointly using the smoothed *F0* contour as the target. The fine-tuning of all parameters was implemented using a weighted representation of the extracted original *F0* contour in the final step.

Figure 1 shows an example of analysis of the sentence "ta1 xi3 huan0 he1 zhong1 guo2 cha2"-"He likes to drink Chinese tea".

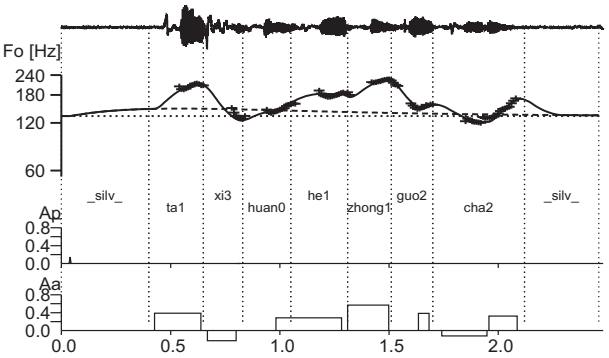

Figure 1: *Speech signal, F0 contour, phrase commands and tone commands of the utterance: "ta1 xi3 huan0 he1 zhong1 guo2 cha2." ("He likes to drink Chinese tea.").*

2.4. Calculation of Tone Command Parameters for Tonal Combination

There are a total of 15 possible combinations of the four tones of Mandarin, as according to a tone Sandhi rule the combination 3+3 is converted to 2+3. The neutral tone does not involve a specific target and was disregarded in this study.

Based on the tone type, tone command configurations for all pairs of syllables were automatically evaluated with respect to the syllabic boundaries, however, not across prosodic phrase boundaries. Tone commands with the same polarity in one syllable were merged. For Tone 2 and Tone 4 the prototypes assume both positive and negative tone commands. If, however, there was only one tone command found within the syllable bounds, the other tone command was searched in the second half of the preceding syllable or the first half of the following syllable, depending on the projected position of the missing tone command.

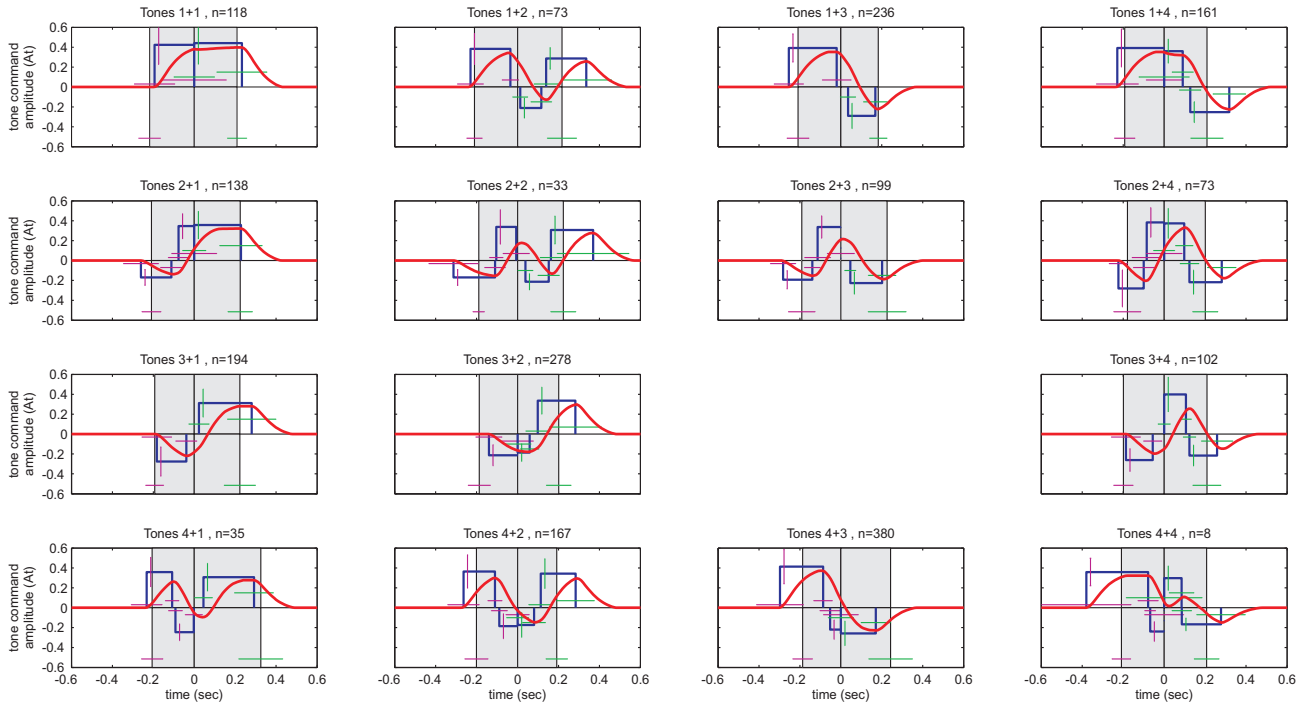

Figure 2: *Means and standard deviations of tone command alignment relative to the syllable onset and tone command amplitude At for all classes of Mandarin tone combination. The vertical line at 0 sec is the boundary between two syllables. The horizontal whiskers indicate the standard deviations of tone command onset time T1 and offset time T2 and syllable offset time relative to the syllable onset time. The vertical whiskers indicate the standard deviation of tone command amplitude At.*

If the missing tone command was not found nearby it was still registered, but with *At*=0. When two tone commands with different polarities were found within the bounds for Tone 1 and Tone 3 syllables whose prototypes project only positive or negative tone commands, the tone command with the right polarity was registered if its duration was greater than 50% of the duration of non-matching tone command.

For every two consecutive syllables with matching tone commands the on- and offset times *T1* and *T2* of tone commands were calculated relative to the respective syllable onsets. Four cases were considered according to the onset and offset of tone commands relative to the syllable onset: a) start before the current syllable and end within the current syllable, b) start and end within the current syllable, c) start within the current syllable and end in the next syllable, d) start before the current syllable and end in the next syllable. The amplitudes of tone commands *At*, the duration of tone commands within the syllable and syllable duration were also calculated. The magnitudes *Ap* of phrase commands were calculated according to the sentence type and speaker gender.

3. Results

The means and standard deviations of on- and offset times *T1* and *T2* which were calculated relative to the syllable onset, amplitude of tone command *At* and duration of every syllable for the 15 cases of tone combinations in Mandarin are shown in Figure 2. It contains also the number of tone combinations for every case. The mean values of alignment and amplitude of tone commands for all tone combinations are represented by the box-shapes. The gray areas indicate the average of syllable duration of the two adjacent syllables. The resulting smoothed *F0* contours for all tone combinations are the responses of the tone control mechanism of the Fujisaki model to the averaged tone commands. The standard deviations for *T1*and *T2* as well as syllable duration are indicated by horizontal lines, and by vertical lines for *At*.

There are cases in which the averaged tone commands start before or end after the boundary between the two syllables examined. Therefore the parts of tone commands which extend beyond the syllable boundary are not displayed and they were also ignored when plotting the resulting *F0* contour. As can be seen from the figure, due to tonal coarticulation and hence depending on the preceding and following tone, *F0* contour for one and the same tone vary considerably. As mentioned above the figure also shows that a sequence of two tone commands with the same polarity can be merged into a single tone command, for instance tone combination 1+1, without significant loss of accuracy because the mean amplitudes on either side of the syllable boundary differ only by a fraction, typically less than 10%. The duration of a tone command is also affected by the polarity of the following tone command.

The tone command amplitude *At* associated with the second syllable is slightly smaller when a tone command of different polarity precedes it (compare the tone commands associated with the second syllable in the columns of Figure 2). In contrast, it was found that there is no significant difference in the *At* of the command associated with the first syllable when the second tone is varied (compare the tone commands assigned to the first syllable in the rows of Figure 2). Figure 2 also shows that even under the influence of coarticulation the characteristic tonal properties of the tones in the adjacent syllables are preserved, that is, for instance, Tone 2 exhibits a rising *F0* contour and Tone 4 a falling one, only the precise timing and amplitudes of the movements vary. If we examine the timing of the averaged tone commands we see that the mean distance between the offset time *T2* of the current command and the onset time *T1* of the following command

with opposite polarity is very small, typically less than 30msec, which is much less than the individual standard deviations for *T1* and *T2*. With respect to the *F0* contour, rising or falling tone switches occur at these locations. In order to simplify the description we therefore propose that *mean* (*T2, T1)* is assumed as the time of the tone switch. Hence, if we assume this modification, the tone command sequence underlying the *F0* pattern is simplified. Figure 3 shows all four combinations for Tone1 on the first syllable applying this simplification. Furthermore, the mean of amplitudes *At* of tone commands of same polarity to the left and right of the syllable boundary was assumed as the amplitude of a merged command crossing the boundary. The smooth red contour indicates the *F0* contour for the non-simplified parameters taken from Figure 2, and the green contour results when both simplifications are applied. It is obvious that the difference is very small, on the average less than 0.3 semitones.

Syllables at the beginning and end of an utterance are special cases as they are not preceded or followed by another syllable. We therefore calculated their properties separately. The duration of tone commands at the end of sentences is greater than at the beginning of sentences for all tones. There was no significant difference concerning the amplitude *At* of tone commands associated with the same tones located either at the beginning or end of utterances with the exception that Tone 3 has higher mean amplitude *At* at the end of an utterance.

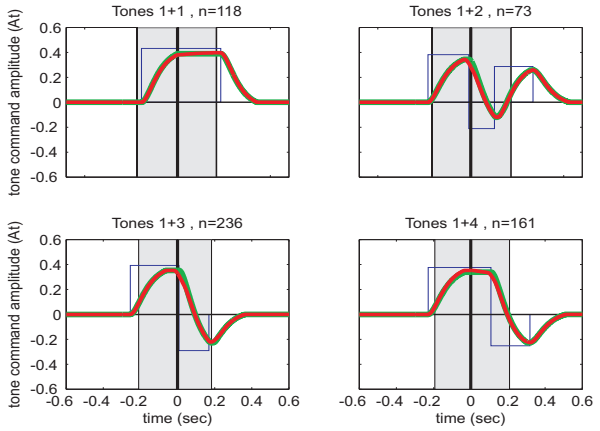

Figure 3: *Simplified tone command parameters for all combinations with leading tone 1. The measured mean tonal contours from Figure 2 (red) and the ones based on simplified parameters (green) are extremely similar.*

The mean amplitude of phrase commands in imperative sentences is greater (*Ap*=.30) than in declarative sentences and questions (*Ap*=.24). The correlation between the number of syllables in an utterance and *Ap* was found to be significant (ρ=.203, p <.01). There are no significant differences in the amplitude of phrase commands between female and male speakers (mean and standard deviation of phrase command amplitude of female speakers are .27 and .19 and for male speakers .21 and .16, respectively). The fact that values are greater than 0 indicates that the phrasal contour must be taken into account when synthesizing *F0* contours of Mandarin.

4. Discussion and Conclusions

This paper examined the effect of tonal coarticulation in Mandarin on the amplitudes and timings of tone commands. To this end, the *F0* contours and Fujisaki-model parameters were extracted from a corpus of short Chinese utterances. In order to develop a tone synthesis model for a Mandarin training system, mean configurations for 15 combinations of two tones were calculated. These indicate that tone commands of the same polarity on either side of syllabic boundaries can be merged. Furthermore, since they pertain to the same tone switch, offset times *T2* and onset times *T1* of neighboring tone commands of opposite polarities can be set to the same value. These simplifications reduce the number of parameters associated with a single syllable considerably. At the most, two timing values of tone switches need to be taken into account, one inside the syllable and one close to the right syllable boundary (combinations 2+2, 2+3, 4+1, 4+4), as well as two tone command amplitude values *At* when commands of alternating polarity occur in the syllable. However, in most cases one of these amplitude values is inherited from the preceding syllable (exceptions are combinations 3+4 and 4+4). The results concerning phrase commands suggest that declination must be taken into account when synthesizing *F0* contours of Mandarin. Based on these results a model for tone synthesis will be developed and perceptually tested.

5. Acknowledgements

This work is funded by the German Ministry of Education and Research grant 1746X08 and supported by DAAD-NSC (Germany/Taiwan) and DAAD-CSC (Germany/China) project related travel grants for 2009/2010.

6. References

[1] Mixdorff, H., Külls, D., Hussein, H., Gong, S., Hu, G. and Wei, S., "Towards a Computer-Aided Pronunciation Training System for German Learners of Mandarin", Proc. of SLaTE 2009, Wroxall Abbey Estate, Warwickshire, England, 2009.

[2] Hussein, H., Mixdorff, H., Do, H. S., Wei, S., Gong, S., Ding, H., Gao, Q. and Hu, G., "Towards a Computer-Aided Pronunciation Training System for German Learners of Mandarin - Prosodic Analysis", Proc. of Workshop on Second Language Studies, Tokyo, Japan. September 2010.

[3] Wang, W. S.-Y., "Phonological Features of Tone", International Journal of American Lingustics, Vol. 33, 2, pp. 93-105, 1967.

[4] Fujisaki, H. and Hirose K., "Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese", Journal of the Acoustical Society of Japan (E), Vol. 5, 4, pp. 233-242, 1984.

[5] Fujisaki, H., Hirose, K., Halle, P., Lei, H. "Analysis and Modeling of Tonal Features in Polysyllabic Words and Sentences of the Standard Chinese", Proc. of ICSLP, pp. 841-844, 1990.

[6] Fujisaki, H., "The Roles of Physiology, Physics and Mathematics in Modeling Prosodic Features of Speech", Proc. of Speech Prosody 2006, Dresden, Germany, May 2006.

[7] Prom-on, S., Xu, Y. and Thipakorn, B. "Modeling Tone and Intonation in Mandarin and English as a Process of Target Approximation", Journal of the Acoustical Society of America, Vol. 125(1), pp. 405-424, 2009.

[8] Mixdorff, H., Fujisaki, H., Chen, G. P. and Hu, Y., "Towards the Automatic Extraction of Fujisaki Model Parameters for Mandarin", Proc. of Eurospeech 2003, Vol. 2, pp. 873-876, Geneva, Switzerland, 2003.

[9] Tseng, C.-Y., Su, Z.-Y., and Lee, L.-S., "Mandarin Spontaneous Narrative Planning—Prosodic Evidence from National Taiwan University Lecture Corpus", Proc. of Interspeech, Brighton, U.K., pp. 2943-2946, 2009.

[10] Wang, R. H., Liu, Q. F. and Wei, S., "Putonghua Proficiency Test and Evaluation", Advances in Chinese Spoken Language Processing, Chapter 18, Springer press, pp. 407-430, 2006.

[11] Boersma, P. and Weenink, D., "Praat doing Phonetics by Computer", version 5.0.42, www.praat.org.

[12] Mixdorff, H., FujiParaEditor ,http://public.bht-berlin.de/~mixdorff/thesis/fujisaki.html, 26.01.2011.