Learning tone distinctions for Mandarin Chinese

David Weenink[1], *Guangqin Chen*[2], *Zongyan Chen*[3], *Stefan de Konink*[4], *Dennis Vierkant*[5]
Eveline van Hagen[6] *& Rob van Son*[1]

[1]Institute of Phonetic Sciences, University of Amsterdam, The Netherlands
`David.Weenink@uva.nl, R.J.J.H.vanSon@uva.nl`
[2]Rotterdam Business School, Hogeschool Rotterdam
`g.chen@hro.nl`
[3]Hogeschool van Amsterdam
`chn@hesasd.nl`
[4]Student, University of Amsterdam
`skinkie@xs4all.nl`
[5]ITBE, University of Twente,
`D.Vierkant@utwente.nl`
[6]Fontys Hogescholen, Eindhoven
`ailin04@yahoo.co.uk`

Abstract

We describe the SpeakGoodChinese system that supports beginning students of Mandarin Chinese to produce tones correctly (http://speakgoodchinese.org/). Students pronounce a word spelled in pinyin notation and receive feedback from our system on their production of the tones. The novelty in our approach lies in the use of synthetic reference tone(s) produced from the pinyin notation. Preliminary results indicate a 6% rejection rate for six words, read multiple times, by three reference speakers and less than 15% acceptance rate on incorrectly produced tones on shadowed versions of these words by 8 speakers. With speech from 4 reference speakers collected with a fully functional test application, a rejection rate of less than 15% was achieved.

Index Terms: Mandarin tone recognition, Dynamic Time Warp (DTW), learning aid

1. Introduction

Mandarin Chinese is the official administrative language of China. It is a tonal language, i.e. a change of tone alters the meaning of a word as much as a change of phonemes. These tones pose major problems for Dutch students and other students with a non-tonal language background who start learning Mandarin. In order to aid these students, we have developed a computer program to assist them in producing the correct tones.

Traditionally a distinction is made of four tones although modern literature also accepts the neutral fifth tone as an additional tone [9]. The most important perceptual attribute of these tones is pitch. In this study we do not consider the interactions between tone, stress and intonation because phonetic studies on the acoustic characteristics of tone have shown that pitch is the primary cue in tone perception [4]. In the literature these pitches are mostly described on a (logarithmic) frequency scale, sometimes with five equidistant tone levels which are named low, halflow, middle, halfhigh and high level. These names facilitate describing the dynamics of the tones [6]. These levels are always relative levels and very much speaker dependent: to identify tones differing only in pitch, listeners must refer to their knowledge of the speaker's pitch range, and where the tone occurs within that range. A low tone produced by a high-pitched speaker like a women and a high tone produced by a low-pitched speaker like a man, may be very similar [4]. In pinyin, the romanized version of the pronunciation of Mandarin, these tones are indicated with tone accents above the vowels. These accents may differ slightly according to circumstances [9, pag 20]. In figure 1 the citation form of the tones of mandarin are given for the word *da*.

The first tone, indicated with a bar above the vowel as in $m\bar{a}$ (mother), starts at the high level and stays that high. The duration of the first tone is somewhat longer than the average duration. Dutch speakers are inclined to speak this tone at too low a frequency.
The second tone, indicated with an acute symbol as in $m\acute{a}$ (hemp), increases in pitch from the middle register to above the level of the first tone. The duration of this tone is somewhat shorter than the average duration and it is perceived as a rather abrupt rise. Dutch speakers have difficulties in making this tone rise enough.
The third tone, indicated with a check symbol as in $m\check{a}$ (horse), starts halflow, falls to low and then (optionally) rises fast to halfhigh. In the low trajectory the voice can easily become creaky. The pronunciation of the third tone is always influenced by the following tone: a third tone preceding a third tone becomes a second tone. A third tone preceding all other tones becomes a half third tone, i.e. only the transition from halflow to low is pronounced and the rising part is not. In [9] more examples of tone sandhi are treated.
The fourth tone, indicated with a grave accent as in $m\grave{a}$ (scold), decreases rapidly from high to low level. Dutch speakers have difficulties in making this tone fall enough.
The (fifth) neutral tone has a short duration and lacks any emphasis. The level of this tone depends on the preceding tone.

The learning aid that we have developed is designed for students that start learning Mandarin and want to receive feedback on their tone productions. Applications that help people

Table 1: The rules for tone simulation. The columns show from left to right the tone, the start, the intermediate and the end frequency of a tone contour, respectively. The value for the intermediate frequency, F_i, is only needed for tone 3 and denotes the lowest frequency. The abbreviation "l.i." means linear interpolation. The symbol δ is 1 semitone. The factor s_f under normal circumstances equals 0.5 and scales frequencies down by one octave. F_l denotes a frequency value derived from the end frequency of the previous syllable of the word.

Tone(s)	F_b	l.i.	F_e
1	F_h	F_h	F_h
2	$F_{b2} = F_h\sqrt{s_f} - \delta$	l.i.	$F_h + 2\delta$
3	$F_{b3} = F_h\sqrt{s_f}$	$F_{i3} = F_h s_f - 3\delta$	F_{b3}
4	$F_{b4} = F_h + 2\delta$	l.i.	$F_{b4} s_f$
0	F_{0b}	l.i.	$F_{0b} - \delta$
6	$F_d = F_h\sqrt{s_f}$	l.i.	$F_d - \delta$
2 1	F_{b2}	l.i.	$F_{b2} + 5\delta$
1 **2**	$F_{b2} - \delta$	l.i.	$F_h + 2\delta$
4 **2**; 3 **2**	$F_l + \delta$	l.i.	F_h
3 1; **3** 4	F_{b3}	$F_h s_f + 2\delta$	F_{b4}
3 2	F_{b3}	$F_{i3} = F_h s_f + 2\delta$	F_{i3}

in realizing correct tones were pioneered by IBM with their speechviewer [8]. Our application starts with a selection of this students tone 1. This frequency, F_h, will be used as a reference frequency for all the other synthesized tones. Then the student hears a pre-recorded word whose pinyin notation is drawn on the screen and he tries to say that word with the correct tone. The pitch of the spoken word will be analyzed. The student will then receive feedback on how well the pitch of the spoken word matches the desired pitch. The next word will be displayed and the cycle starts again.

The novel aspect of this learning aid is that we do not need sub-standard pre-recorded words for a proper tracing of incorrect student utterances. Based on the results of preliminary inquiries with students and teachers, our recognizer was biased towards accepting student utterances to prevent de-motivating them.

We use an intelligent algorithm to calculate a number of possible erroneous pitch tracks from the pinyin word. A DTW-algorithm was implemented in the PRAAT program [1] that aligns the pitch of the spoken word with pitches of all possible combinations of pitches and picks the best match. The selection of the best fit was biased to the correct tone combinations by allowing the correct choice to be within a short distance of the best match. A few post-matching rules implement some assimilation phenomena. We use six different tones and our words are maximally two syllables long: we therefore need to consider 29 possibilities (the first syllable cannot carry a neutral tone and the *3-3* and *2-3* tone combinations are identical). The sixth tone is the tone that would correspond to standard Dutch intonation and bears no resemblance to any of the Mandarin tones. A decision is made on the quality of the pitch of the spoken word in comparison to all the synthesized pitches, and proper feedback is given.

2. Pitch synthesis

The words we use in this study consist of only one or two syllables. These syllables can be easily identified from the pinyin notation because here pitch is indicated by a number after each syllable. For example, the first syllable in *duo1shao3* has tone 1 and the second syllable has tone 3. For the synthesis of the *1&3* tone contour of this word we must synthesize the correct duration and the correct tone contours.

Duration is calculated with a heuristic that is based on the consonants in the obstruent parts and the number of consonants and vowels in the sonorant parts. The starting sound of a syllable is either a sonorant or a non-sonorant. A word may be totally sonorant or or it may consist of two sonorant parts separated by an unvoiced part. The duration of the obstruent part is fixed at a value $d_{uv} = s_t \times d_s$, where the tone scalefactor s_t has the value 1, 0.8, 1.1, and 0.8 for tones *1* to *4*, respectively, and d_s is an average segment length of 0.15 s. The sonorant parts have duration $d_v = s_t(d_s \times n_v + d_f)$, where n_v is the number of segments in the sonorant part, with di- and triphones counting as 2 or 3 segments, and d_f is a fixed duration of 0.12 s. The tones are completely realized on the sonorant parts of the syllables.

The parametrization of the generated pitch contours all depend on F_h, which is also the value for tone *1*. The value for F_h was determined by adjusting a global, gender specific value, to get a best fit of the range of the correct tone contour on the pitch track of the spoken word. Adjustments were limited to 3 semitones downwards and 6 semitones upward from either 200 Hz (men) or 300 Hz (women). The pitch excursion size was standardized to get a 1 octave (12 semitones) excursion size for the fourth tone. This was fitted on the word spoken by the student, limited to excursion values between 9-18 semitones scaled to tone *4*. The bias towards higher pitch and excursion sizes was motivated by the natural tendency of students to exaggerate tone pronunciation. For the synthesis of the other tones some rules of tone sandhi have been incorporated too.

Table 1 shows for each single tone, indicated by column header Tone(s), what the values are for the start frequency and the end frequency of the contour. These frequencies are indicated by the column headers F_b and F_e, respectively. F_i, an intermediate frequency is only relevant for tone *3* and indicates the lowest frequency that has to be synthesized for the tone contour. If no extra frequency scaling is involved, the frequency scale factor s_f equals 0.5 and therefore lowers frequencies by one octave. The factor $\sqrt{s_f}$ then signals a lowering by 6 semitones. The start frequency of tone *2* is approximately *1* semitone below $0.7F_h$ and rises approximately linear at a level two semitones above F_h. For tone *3* we start half an octave, i.e. six semitones, below F_h and then lower the frequency until it is fifteen semitones, i.e. one octave plus three semitones, below F_h. This intermediate frequency F_{i3} will be the lowest frequency

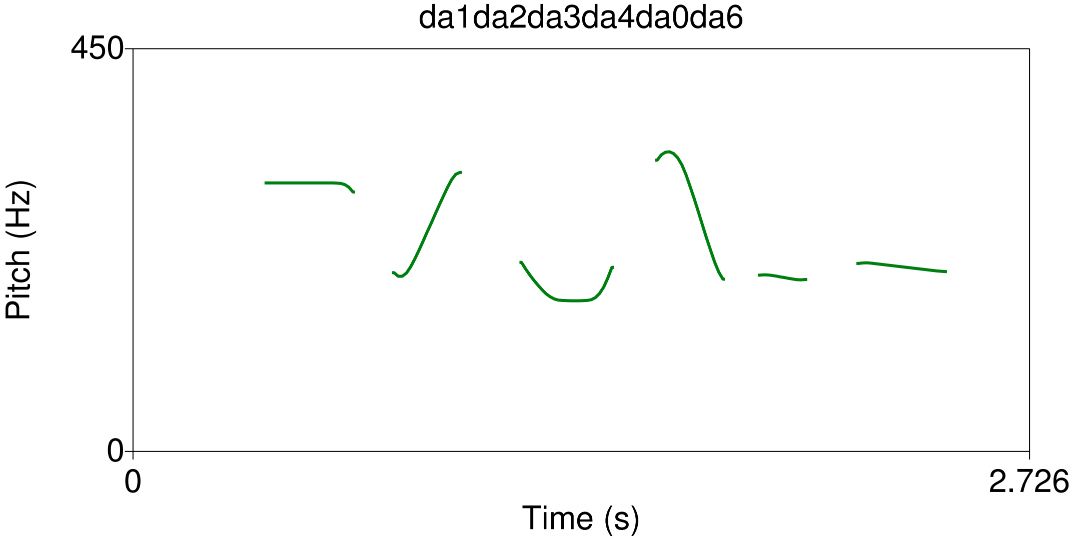

Figure 1: The six tones synthesized for the pinyin word *da*. From left to right the first four standard tones. Then the neutral tone and the last one is the standard Dutch intonation tone.

attainable by any tone contour. The start and end frequencies of tone *0*, the neutral tone, are variable. If the end frequency of the previous syllable, F_l, is larger than $\sqrt(s_f)0.7F_h$ the start frequency F_{0b} equals $F_l - 3\delta$. Otherwise it is set to $F_l + 3\delta$. For the end frequency a 1 semitone declination is subtracted. Tone *6* is special because it does not correspond to any Mandarin tone. It is only present to be able to give better feedback for cases where the student did not realize any Mandarin tone but instead fell back on Dutch intonation. The Dutch intonation starts 6 semitones below F_h and falls by one semitone.

The second part of the table shows the effects of tone sandhi. The tone undergoing this change is indicated in bold typeface. In the first line with "**2** 1", we see how tone *2* changes under the influence of a following tone *1*: its excursion size is reduced. The next line starts with "1 **2**" and shows how tone *2* changes under the influence of a preceding tone *1*: its excursion size is expanded by one semitone.

These pitch contours do not capture all the natural contextual variation. Especially the realizations of the tone 2 and 3 in context were often ambiguous. But we also observed neutralization of tone movements in the other tones. Therefore, rules were added that merged bi-syllabic tone combinations into super classes. These rules reduced the tone perplexity at the cost of increased false acceptances.

3. Pitch measurements

The studies considered here [2, 3, 4, 6, 7] do not mention the problems that might occur in the measurement of the pitch of Mandarin tones. For example, in the fast rising pitch in preparation of a tone *1*, or in the fast falling pitch for tone *4*, large pitch changes occur in a very short time which means that only short measuring intervals can be used. During the production of the very low pitches for the third tone voicing may become creaky, creating a discontinuity in the measured pitch pattern. The pitches of the spoken words were measured with the PRAAT program. In order to cope with possible local pitch measurement errors we use a Dynamic Time Warp (DTW) algorithm for matching pitch contours. Furthermore for correctly assigning correct tones, the pitch range of the speaker had to be established adaptively as the pitch height and ranges of individual words differed too much.

4. Results

The tone recognizer was tested for three different test sets. The subjects were 5 reference speakers for correct pronunciation, four of which are coauthors of this paper: 3 native speakers of Mandarin and 2 very advanced native Dutch speakers. With respect to these tests, we did not find a difference between the native speakers and the advanced Dutch speakers. Incorrect pronunciations were obtained from 3 of the reference speakers and 5 Dutch students with (very) low proficiency. The following tests were done with the March 9, 2007 version of the recognition engine.

1. Examples of correctly pronounced tokens were collected using a Tcl-based prototype user interface. Four reference speakers typed mono- and bisyllabic pinyin words of their choosing and pronounced them. From a total of 399 collected tokens, 57 were rejected (14%).
2. The six pinyin words *cha2*, *gong1zuo4*, *jie2hun1*, *dian4hua4*, *duo1shao3* and *shi2jian1* were read several times by three reference speaker. The tone recognizer rejected 5 out of 83 tokens (6%).
3. The previous six words where shadowed from recorded examples with incorrect tone realization by 8 low and high proficiency speakers. The most obvious problem with incorrect tone pronunciation is a very low pitch and small excursion size and our recognizer automatically flags this as an error. Therefore, all tokens with a maximum pitch or excursion size 3 semitones less than the reference tone contour were rejected. In total 37 of the 320 words were accepted (12%). Most of the accepted ones were due to incorrect classification of the second tone in the *cha2* pinyin word.

Overall this shows a false rejection/acception rate of less than 15% with ample opportunities for improvement.

5. Conclusions

We have developed a real-time tone recognizer with an 85% acceptance rate for DTW pitch contours of Mandarin Chinese. The method is biased for acceptance of one- and two syllable words. Besides tone, it is possible to extend the method to also test the pronunciation of these words. The determination of wrongly produced tones is not optimal yet but can be improved upon.

6. Acknowledgements

The SpeakGoodChinese project was made possible by a grant from the Digital University.

7. References

[1] Boersma, P. & D. Weenink, Praat: doing phonetics by computer (Version 4.5.15), Computer program. URL: http://www.praat.org/, 2007.

[2] Dong, M., H. Li & T. L. Nwe, ”Evaluating prosody of Mandarin speech for language learning”, in *Proc. Interspeech*, 1986–1989, 2006.

[3] He, L. & J. Hao, ”A tone recognition framework for continuous Mandarin speech”, in *Proc. Interspeech*, 1575–1578, 2006.

[4] Jongman, A., Y. Wang, C. B. Moore & J. A. Sereno, ”Perception and production of Mandarin Chinese tones”, in *Handbook of Chinese Psycholinguistics*, Cambridge Universiy Press, 2006.

[5] Mixdorff, H. & Y. Hu, ”Word structure and tone perception in Mandarin”, in *Proc. Interspeech*, 1507–1510, 2006.

[6] Peabody, M., S. Seneff & C. Wang, ”Mandarin tone acquisition through typed interactions”, in *InSTIL/ICALL Symposium: NLP and Speech Technologies in Advanced Language Learning Systems*, 2004.

[7] Sun, Y., D. Willet, R. Brueckner, R. Gruhn & D. Bühler, ”Experiments on Chinese speech recognition with tonal models and pitch estimation using the Mandarin Speecon data”, in *Proc. Interspeech*, 1245–1248, 2006.

[8] Wempe, T & M. van Nunen, ”The IBM speechviewer”, in Proceedings IFA **15**, University of Amsterdam, 121–129, 1991.

[9] Wiedenhof, J., Grammatica van het Mandarijn, Uitgeverij Bulaaq, Amsterdam, 2004.

[10] Xu, Y. & X. Sun, ”Maximum speed of pitch change and how it may relate to speech”, J. Acoust. Soc. Am. **111**, 1399–1413, 2003.