Chinese Tone Modeling with Stem-ML

Chilin Shih and Greg P. Kochanski

Bell Labs – Lucent Technologies, Murray Hill, NJ 07974, USA

ABSTRACT

This paper models tonal variations with Stem-ML tags. Surface tone shapes often deviate from their expected canonical shapes in natural sentences, presenting a challenging case to tone modeling. In this study we employed a subset of Stem-ML tags which incorporated information of lexical tones and linguistically motivated prosodic strength of the syllable. The tags successfully captured the "distorted" tone shapes and produced contextually appropriate surface variations.

1. Introduction

This paper uses Chinese data to test the ability of Stem-ML to model tone/accent interaction. The Stem-ML tag set is described in a companion paper [1], and at http://www.bell-labs.com/project/tts/stem.html.

When tones or accents occur in a crowded space, their interaction with their neighbors leads to surface variations. Analogously, tones are treated in Stem-ML as soft templates which are subject to modifications by other constraints, such as neighboring tones and the phrase curve.

Chinese has several properties that are suitable for testing tonal interactions. The tone inventory of Chinese is known, which takes the guess work out of how many tonal categories there are, what tonal templates look like, and which template to use. Chinese tones come in close proximity with each other with very few tonal co-occurrence constraints. These properties create ideal experimental conditions where interesting interactions occur frequently, and we can strategically utilize the known factors (tonal categories) to explore the unknown (tonal interactions). Conversely, better understanding of tone or accent interaction will generalize to languages where the category information is not lexically given.

2. Chinese tones and their variations

Mandarin has four lexical tones traditionally named tone 1 to 4. Their shapes are high level, rising, low (with optional rising tail in the sentence final position) and falling, respectively. The domain of the tone is the syllable, although we show that tones affect each other in nearby syllables. The four tone shapes are graphed in Figure 1. The tonal contours were calculated from a database of female speech reading all possible Chinese syllables embedded in a frame. Each syllable shows characteristic segmental effects which add variation to the tone shape. We ran a linear regression analysis using five factors to fit the observed pitch values: tones, initial consonants, glides, vowels, syllable codas. The resulting tone coefficients were plotted in Figure 1. This procedure was used to minimize segmental effects. It turned out that the result matched very closely the average curve of each tone.

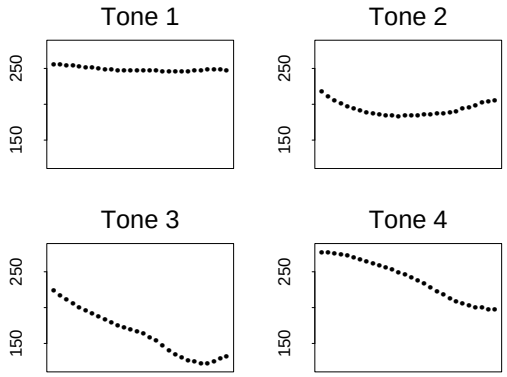

Figure 1: Four lexical tones of Mandarin Chinese

It is possible for a prosodically weak syllable to be toneless, a phenomenon being traditionally termed *neutral tone*. The pitch contours of the neutral tone syllable is conditioned primarily by the tone of the preceding syllable, although other factors such as the following syllable also play a role.

Production studies of Chinese tones show that tone shapes often deviate from the expected canonical shape in natural sentences. The situation is particularly difficult in conversation where the boundaries among tonal categories are blurred. In extreme cases, a tone may be realized with a shape opposite to the lexical specification. We show several examples below, all taken from conversational data.

Figure 2 displays a mild case of tonal distortion. This figure shows the pitch track of *mu3 dan1 huar1* "peony". The second syllable *dan1* and the final syllable *huar1* both have lexical tone 1, the high level tone. While the pitch contour of the final syllable maintains the characteristic level shape, the pitch contour of the middle syllable, in contrast, appears to be rising rather than level. This distortion can be explained by its adjacency to the preceding low tone. The degree of distortion is not severe enough to impact the correct identification of the tone. When excised out of context, the syllable *dan1* was still identified by native listeners as tone 1.

In Figure 3 *fan3 ying4 su4 du4* "reaction time", the second, third and fourth syllables all have lexical falling tone. While the third and fourth syllables *su4 du4* maintain the falling shapes, the second syllable *ying4* is unexpectedly rising. The tonal trajectory is the opposite of what we expect from the lexical tone. The distortion can be explained by the environment: the immediately preceding pitch is low, therefore the beginning pitch of *ying4* is low; the immediately following pitch is high, hence the end of *ying4*

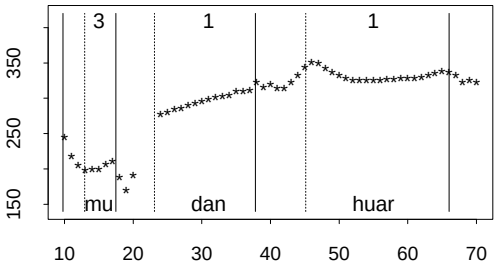

Figure 2: Pitch track of *mu3 dan1 huar1*, where the second syllable *dan1* carries a distorted high level tone, with a surface rising shape that reflects the influence of the preceding low tone.

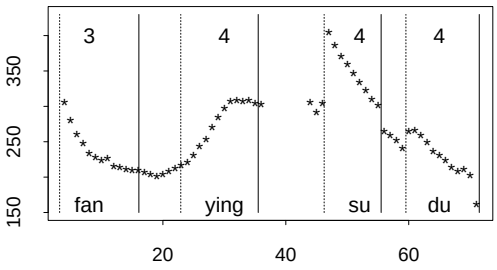

Figure 3: Pitch track of *fan3 ying4 su4 du4*, where the second syllable *ying4* has a rising pitch contour, instead of the expected falling one.

is high. When excised out of context, *ying4* was consistently perceived as tone 2.

Figure 4 *mai3 mai4 jiu4 che1* "buy and sell old cars" mirrors the earlier example, and the same distortion occurs. In this figure, the second syllable is again rising despite its lexical falling tone specification, and again the surrounding contexts are in direct conflict with the tonal specification. Also, the pitch contour of *che1*, the final syllable of the word, is pulled down by the low beginning of the following word *wo2* "I" which has an underlying tone 3 but is changed to tone2 by a phonological process. When excised out of context, *mai4* was perceived as either tone 2 or tone 1.

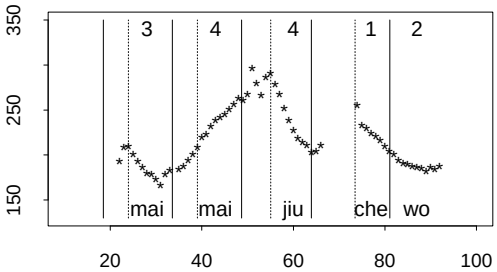

Figure 4: Pitch track of *mai3 mai4 jiu4 che4*, where the second syllable *mai4* has a rising pitch contour, instead of the expected falling one.

The tonal variations cited above from conversational data are consistent with production experiments [4, 3] where unexpected tone shape is associated with weak prosodic strength, and a weak tone accommodates the shapes of neighboring strong tones. The effect is further verified by perception experiments [7]. Xu classified tonal environments as *friendly* and *unfriendly*. Friendly environments refer to the ones where adjacent tonal specifications are the same, such as when a high level tone is connected with another high level tone on either side, or is followed by a high falling tone. Unfriendly environments refer to the ones where adjacent tonal specifications are different, such as when a rising tone (which starts low) follows a rising tone or a high level tone (which ends high). In production, tones in *friendly* environments maintain their specified tone shapes better than tones in *unfriendly* environments. In perception, excised tones from *friendly* environments have higher chances of being identified correctly, compared to tones excised from *unfriendly* environments. Both production and perception data suggest that, if the intended tonal trajectory of a weak tone is in contradiction to the tonal specification of adjacent strong tones, the weak tone gives in. In the two previous examples, the distorted tone shapes both occur on weak syllables (the second syllables of multi-syllable words), and the observed distortion conforms with the neighbor's influence.

The phenomenon of tonal variation is known in the literature. But efforts in TTS tone modeling lagged behind in this domain. In many Chinese TTS systems [2, 5, 6], the TTS generated tone shapes match the lexical expectation, namely, a lexical rising tone will be given a rising pitch contour and a falling tone a falling pitch contour. Neglect of tonal variation results in over-articulation of tones that contributes to the unnatural speech that characterizes many TTS systems.

In the following, we model the phenomenon of tonal variations with Stem-ML. Tones or accents of a language are represented as soft templates in Stem-ML, which bend to conform to their environment. The elements of the environment that affect the tone shapes include, minimally, the preceding tones, the following tones, and possibly higher order phrasal effects that affect a larger region than the tone. We assume one strong constraint in Stem-ML, that is the transition between tones must be smooth. When there is a conflict, Stem-ML uses weights to control how the smoothness constraints is satisfied. Different degrees of tonal variation can thus be approximated by varying weights.

3. Experiment Design

We tested the properties of Stem-ML tags using experimental sentences of Mandarin Chinese which are rich in controlled tonal and prosodic strength variations.

The sentences have the general form of

X	duo1	ying1-gai1	deng1	bi3-jiao4	duo1
X	more	should	lamp	comparatively	more

"If there is more X, then there should be more lamp."

The keyword X is either a monosyllabic word *yan* with any of the four lexical tones, or a tri-syllabic word where the middle syllable *yin* or *ying* has one of four possible tones. The eight possible key-

words are given below. The starred words *camcorder* and *projector* were coined to provide near minimal sets for the experiment. The coined expressions conform to Chinese word formation practice and subjects readily accepted these terms.

煙	鹽	眼	燕
yan1	yan2	yan3	yan4
smoke	salt	eye	swallow
收音機	**收銀機**	**收影機**	**收映機**
shou yin1 ji	shou yin2 ji	shou ying3 ji	shou ying4 ji
radio	cash register	camcorder*	projector*

Alternating tone 1 to tone 4 in the keywords provides experimental control of tonal variations. The contrast of monosyllabic and trisyllabic words provides experimental control of variations in prosodic strength. Figure 5 depicts the typical prosodic pattern of these two types of words in Chinese. A monosyllabic word spoken in isolation is strong (S), as shown in the non-branching prosodic tree to the left. A trisyllabic word has a more complicated rhythmic pattern with alternating strong (S) and weak (W) syllables, as shown in the tree to the right. At least theoretically [8, 9, 4], the test syllable *yan*, being a monosyllabic word, is in a strong position while the test syllable *yin* and *ying*, being the second syllable of a trisyllabic word, is in a weak position.

4. Modeling

We generated F0 contours automatically using Stem-ML tags. To do this we need three types of information: (1) Soft templates for lexical tones, (2) prosodic strength specification, (3) characteristic phrasal contour of the speaker or the style.

We used four soft templates for the four lexical tones. The templates described the tone curves shown in Figure 1 with six equal-distant samples. They were obtained from a database of female speech and were used to model male speech in this study. The pitch range was normalized by taking the highest value of tone 4 (the second value) as 1, and the lowest value of tone2 (the third value) as 0.

Normalized Time	0%	20%	40%	60%	80%	100%
Tone 1	0.63	0.76	0.65	0.64	0.66	0.68
Tone 2	0.34	0.26	0.00	0.02	0.12	0.26
Tone 3	0.23	0.16	-0.32	-0.57	-0.41	-0.21
Tone 4	0.85	1.00	0.71	0.47	0.25	0.18

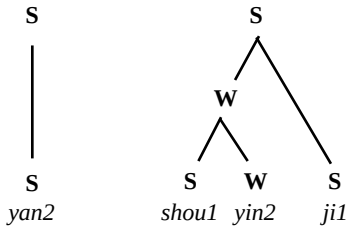

Figure 5: Prosodic strength of Mandarin monosyllabic and trisyllabic words.

Chinese	English	Strength	Type
shou-	radio	1.5	0.5
yin-	–	1.0	0.2
ji	–	1.0	0.3
duo	more	1.1	0.5
ying-	should	0.8	0.2
gai	–	0.8	0.3
deng	lamp	1.0	0.5
bi-	comparatively	1.5	0.5
jiao	–	1.00	0.3
duo	more	1.00	0.5
yan	smoke	1.5	0.5
tone3		1.5	0.5

Table 1: The Stem-ML stress tags used to generate F_0

The tone shapes are controlled by Stem-ML *stress* tags, which specify tone templates, strength, and type (see [1] for details). The strength and type values were fitted manually to one sentence with the keyword *shou1 yin1 ji1* in the training set and Figures 6 and 7 are sentences from the test set. Tag values of the monosyllabic *yan* sentences were derived from the trisyllabic *yin* sentence by removing two weak syllables. The strength contrast of *yan* vs. *yin/ying* syllables reflects the difference in strength depicted in the prosodic structure of Mandarin monosyllabic vs. trisyllabic words. Some areas reflect paralinguistic function that occurs in natual speech. For example, there is a discrepancy between the relative strength of *shou*/*ji* and the theory (Figure 5), as a result of iambic reversal. Furthermore, the modal *ying-gai* "should" is weaker than its neighbors.

After the initial fitting, we realized that Tone 3 is more stable than our initial model allows. Implementing Tone 3 with a constant strength and type rather than varying the degree of strength by word prosody gives better results. Figures 6 and 7 show the result of constant Tone 3 specification.

Finally, we fit all sentences in the experiment with the same pitch range specification and the same exponential decay rate of the phrase curve. The pitch range is set to 80Hz–130 Hz, and the phrase curve droops from 105 Hz to 80 Hz in half a second.

Figures 6 and 7 show the predicted F_0 as stars and the observed F_0 in small filled circles. Syllable boundaries are marked by thick vertical lines, and syllable internal consonant/rhyme boundaries are marked by thin dashed lines. The numbers above the pitch tracks indicate lexical tones of the syllables.

The generated F0 fits the observed data very well, particularly in areas where tonal variation is large and traditional analysis (such as target specification) fails. First, there is the contrast between the keywords in Figure 6 and Figure 7. In general, the *yan* syllables have more discernible pitch contours than the contrasting *yin/ying* syllables, as a result of the stronger Stem-ML strength and type specification (1.5/0.5 for yan and 1.0/0.2 for yin). The tonal gestures of *yin/ying* are weak in both the observed and the fitted F_0 of the sentences, with the fitted F_0 erring on the cautious side by retaining slightly more of the original tonal characteristics.

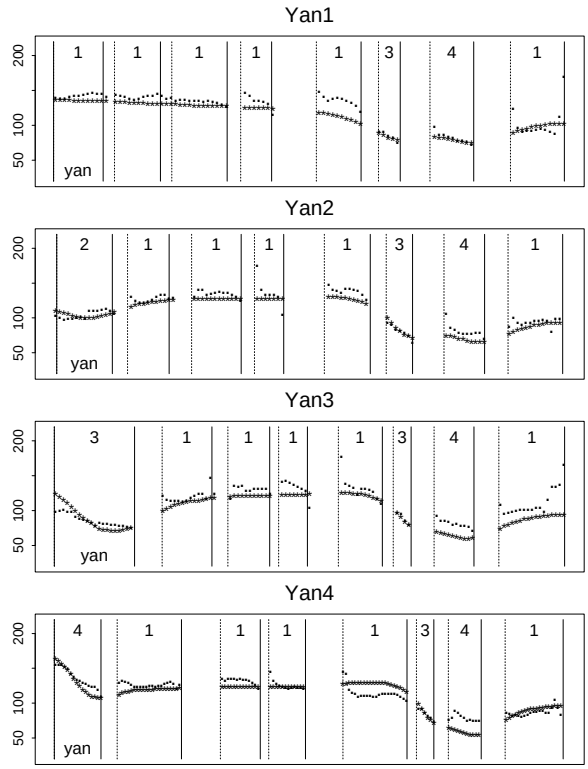

Figure 6: Pitch curves generated by Stem-ML tags – Monosyllabic keywords.

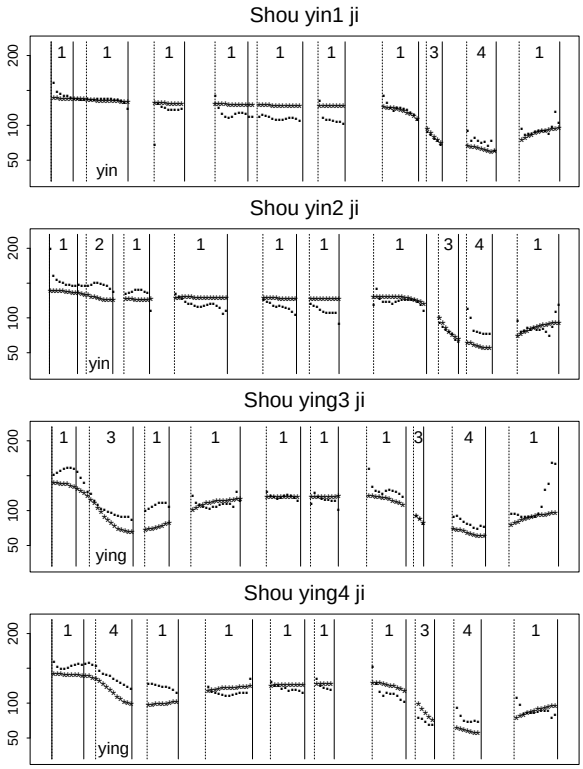

Figure 7: Pitch curves generated by Stem-ML tags – Trisyllabic keywords.

Weak high level tones (tone 1) after low tones or low pitch are rising rather than being level. There are several examples of this, most notably in the syllable after *yan3*, in the third syllable *ji1* of *Shou1 ying3 ji1*, and the final syllables of all the sentences.

Finally, there are several variations of tone 4. The strongest realization is on the syllable *yan4*, the keyword in the last sentence in Figure 6. The next level is realized on the syllable *ying4*, the second syllable of the last sentence in Figure 7. Furthermore, every sentence has a weak tone 4 *jiao4* in the penultimate position. Note that the strength/type specification of *jiao4* is comparable to that of *ying4*, however, the falling shape of *jiao4* is further compromised by the preceding tone3.

5. Conclusion

This paper models tonal variations with Stem-ML tags. Tonal distortion is predictable and can be modeled by varying the relative strength of neighboring tones.

One of the fundamental assumption of Stem-ML is that articulatory gestures, including the control of F_0, are smooth. The conflict is resolved by compromising between all constraints. The instruction of how to weigh the constraints comes from linguistic knowledge, such as the relative strength of words and syllables.

We expect the physical constraints of smoothness to generalize to other languages, and the finding on how to control tonal interaction in a crowded space as in Chinese will help us learn the accent variations of other languages.

6. REFERENCES

1. Kochanski, G. P., and Shih, C. Stem-ML: Language independent prosody description. In *ICSLP 2000* (Beijing, China, 2000).
2. Lee, L.-S., Tseng, C.-Y., and Ouh-young, M. The synthesis rules in a Chinese text-to-speech system. *IEEE Transactions on Acoustics, Speech, and Signal Processing 37*, 9 1989, 1309–1320.
3. Liao, R. *Pitch Contour Formation in Mandarin Chinese*. PhD thesis, Ohio State University, 1994.
4. Shih, C., and Sproat, R. Variations of the Mandarin rising tone. In *Proceedings of the IRCS Workshop on Prosody in Natural Speech* (1992), University of Pennsylvania, pp. 193–200.
5. Shih, C., and Sproat, R. Issues in text-to-speech conversion for mandarin. *Computational Linguistics and Chinese Language Processing 1*, 1 1996, 37–86.
6. Sproat, R. W., Ed. *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*. Kluwer, Dordrecht, 1998.
7. Xu, Y. *Contextual Tonal Variation in Mandarin Chinese*. PhD thesis, The University of Connecticut, 1993.
8. Yip, M. *The tonal phonology of Chinese*. PhD thesis, MIT, 1980.
9. Zhang, Z.-S. *Tone and tone sandhi in Chinese*. PhD thesis, Ohio State University, 1988.