Exploring Tonal Variations via Context-Dependent Tone Models

Yue-Ning Hu [1], *Min Chu* [2], *Chao Huang*[2], *Yan-Ning Zhang* [1]

[1] School of Computer Science and Engineering,
Northwestern Polytechnical University, Xi'an, 710072, China
[2] Microsoft Research Asia, Beijing, 100080, China

ynhu.moon@gmail.com, {minchu,chaoh}@micrososft.com, ynzhang@nwpu.edu.cn

10.21437/Interspeech.2007-360

Abstract

In this paper, we study tonal variations by training context-dependent tone models from a large speech corpus. Each model represents a tone-in-context and can output a stylized f_0 pattern for it. With these tone models, it becomes tangible to investigate f_0-variations with plenty of factors. Six contextual factors are investigated to describe tones in this work. We find that impact of a factor varies across tones as well as the three states of a tone. Normally, onset pitch of a tone is determined jointly by syllable position and left tone, while, the offset pitch is mainly determined by syllable position. For a neutral tone, its pitch level is mainly depended on the left tone and syllable position affects its offset pitch. Both current vowel and right tone influence the pitch level, yet their impacts are weaker than syllable position and left tone, except for T3, the low tone.

Index Terms: tonal variation, context-dependent, tone model

1. Introduction

For a tone language like Mandarin, it is well known that f_0 contours of tones vary extensively in continuous speech. Many works were done to explore why and how such variations happened. Wu sketched the tone-sandi patterns within polysyllabic words [1, 2]. Xu explicitly showed the carry-over effects of the offset of preceding tone on the f_0 contours of four Mandarin lexical tones [3]. Shen described the Mandarin intonation as a series of pitch ranges that are bounded by an upper-line and a bottom-line [4]. The tone-sandi patterns and the two-line intonation model have been deployed in a Mandarin text-to-speech system [5] that demonstrates rather nature phrase intonation. All these studies revealed that f_0 realization of Mandarin tones was affected by many factors. However, when the number of factors considered increases, it becomes very difficult to study the tonal variation caused by these factors (independently or jointly) through the traditional laboratory speech corpus. In this paper, we propose to explore tonal variations through *context-dependent tone models* (CDTM) trained from large scale speech corpora. Each model represents a type of context of a tone and we can derive a stylized f_0 pattern for the *tone-in-context* (TIC). With these CDTMs, it becomes tangible to investigate f_0-variations with plenty of factors.

This paper is organized as follows. In Section 2, contextual factors, training and evaluation procedure of the CDTMs, as well as the results are introduced. Analyses of tonal variations caused by contextual factors are presented in Section 3. Conclusions and a final discussion are in Section 4.

2. Context-Dependent Tone Model

Context-dependent phone models, such as tri-phone Hidden Markov Models (HMM), have been successfully used in speech recognition. However, for Chinese speech recognition, only considering neighbored phones seems not enough to capture tonal variations. In [6], supra-tone models that are trained from out-line features of two or three succeeding tones were used to re-score the tonal-syllable lattice generated with tri-phone models, and achieved significant improvement in tonal-syllable accuracy. In this paper, we take more factors into consider and train CDTMs to depict the relationship between tonal variations and contextual factors.

Our focus is on tones. The tone to be modeled is referred as *current tone*, CT in short. Since the tonal information of a syllable is mainly carried by its final part, CDTMs are built from syllable finals.

2.1. Contextual factors

Many factors may contribute to the variations in tone realization. Some are linguistic/paralinguistic demands, such as prosodic structure, syntax, pragmatics and emotions [7]. Due to the difficulties in generating syntactic, pragmatic and emotional labels for a large speech corpus, in this paper, only prosodic structure is considered and it is described as *syllable position* (denoted as SP) within a phrase as shown in Table 1.

Table 1. *The description of syllable position*

SP1	*phrase initial*, first syllable of a phrase
SP2	*word initial*, first syllable of a prosodic word, but not at phrase initial
SP3	*word middle*, syllable within a prosodic word
SP4	*word final*, last syllable of a prosodic word, but not at phrase final
SP5	*phrase final*, last syllable of a phrase
SP6	*mono-syllabic prosodic word*, it can be at any position within a phrase

Other factors are articulatory constraints, such as vowel intrinsic pitch, f_0 perturbation by consonants, and maximum speech of pitch change and pitch direction shift [7]. In this paper, tones of preceding and succeeding syllables (referred as *left tone*, LT in short, and *right tone*, RT in short, respectively) are considered to capture the carry-over and anticipatory effects between tones. The final of the focused syllable (referred as *current final*, CF in short), the initial of the focused syllable (referred the *left consonant* of tone model, LC in short) and the initial of the next syllable (referred as the *right consonant*, RC in short) are considered to capture the impact of phonetic context.

All together, 6 factors are investigated in this work.

2.2. CDTM training

The training of CDTM is rather similar to that of traditional tri-phone HMMs except that (1) more contextual factors are

considered; (2) all models with the same tone identity are tied into the same tree in the decision-tree based context clustering to capture more tonal variations instead of segmental variations.

2.3. CDTM evaluation

The quality of CDTMs is evaluated in a second-pass tone-recognition task. In the first-pass, syllable-loop decoding is performed with tri-phone HMMs and phone boundaries are generated. In the second pass, each syllable final in the decoded results is rescored by CDTMs of five tones in different contexts. Since the base syllable error rate of the first-pass decoding is only 3.8%, we keep the base syllables unchanged and search for a more suitable tone for each of them in the second-pass. Therefore, the CF, LC, and RC are fixed for each final. Other three factors (SP, LT and RT) are enumerated for five tones. Then, up to 900 CDTM candidates are generated for each final, and each model outputs a likelihood of the acoustic observations of the final. Viterbi search is used to find the best CDTM path that complies with constraints between succeeding models, and tones of CDTMs on the best path are output as the final results.

Here are some examples of the constraints: (1) the CT and the LT of a CDTM should be the same as the RT and the CT of the CDTM before it, respectively; (2) SP1 is only preceded by SP5 or SP6, etc.

2.4. Experiments and Results

2.4.1. Speech corpus and setups

A large Mandarin speech corpus read by a broadcast woman is used in our experiments. It is designed for speech synthesis and covers rich phonetic and prosodic variations. Prosodic structure [8] has been labeled manually, from which SP labels are derived automatically. The corpus contains 14476 utterances, 13476 of which are used for training and the remaining 1000 for testing. Parameters are tuned with the first 100 utterances in the testing set.

Multi-space distribution (MSD) HMM, proposed by Tokuda *et. al.* [9], can model both discrete and continuous features at the same time. When it is used to model pitch patterns, neither of voice/unvoice decision is needed. MSD-HMM has shown benefit to tone recognition in Chinese [10]. In this work, MSD-HMM is used for both tri-phone model and CDTM.

In our experiments, each model contains three left-to-right states. *Minimum description length* (MDL) criterion is used to control the growing of decision trees when models are clustered.

Acoustic features include 39-dimensional MFCC and 5-dimensional f_0 related features, which are logarithmic f_0, its first and second order derivatives, pitch duration and long-span pitch [11].

Base syllable *error rate* (ERR) in the first-pass result is only 3.8% and tonal syllable ERR is 11.7%.

2.4.2. CDTM vs. tri-phone model

The tone ERRs of the first-pass decoding and the second-pass rescoring are all 8.7%. In other words, the second-pass rescoring hasn't achieved any improvement on the tone recognition task. Yet, at least, we can conclude that the CDTMs are not worse than tri-phone models in modeling tones. The benefit of CDTM is that it shows us a clearer picture on how tone varies in different contexts. Therefore, it is a powerful tool for studying tonal variations.

2.5. Generating stylized f_0 pattern with CDTM

A CDTM represents an f_0 pattern of a tone in the given context. Its three states represent the onset, main and offset parts of the tone-in-context, respectively. By plotting the f_0 means of the three states, we can visualize f_0 patterns of a CDTM and study the variations among tones in different context. Figure 1 shows an example. The stylized f_0 patterns of two Chinese sentences are plotted in the same figure to show the f_0 differences caused by prosodic structure and tone. When the first syllable /ta1/ is uttered as a monosyllabic word, its f_0 is lower than when it is grouped into the same prosodic word as the next word /yi3 jing1/. The two patterns at the fifth syllable show the f_0 patterns of T2 (for syllable /shi2/) and T1 (for syllable /gong1/) in a similar context, respectively. In sent2, three syllables (the first, /ta1/, the third, /jing1/ and the fifth ones, /gong1/) carry T1. Their pitch levels decreases from phrase initial toward phrase final.

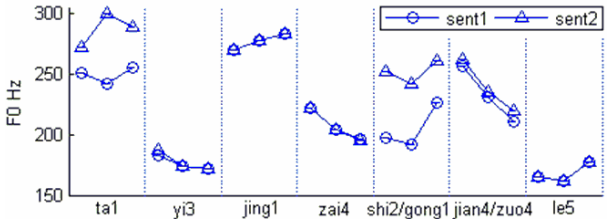

Figure 1. Stylized f_0 pattern of two Chinese sentences (Sent1: 他已经在实践了 |ta1# yi3 jing1# zai4# shi2 jia4 le5; Sent2:他已经在工作了|ta1 yi3 jing1# zai4 # gong1 zuo4 le5; # marks the prosodic word boundaries, number 1-4 mark the four Mandarin tones and 5 represents the neutral tone).

Table 2. *Confusion matrix of tones*

ref / % / rec	1		2		3		4		5	
	4062		5034		3010		6739		1208	
	Mfcc+f_0	f_0	Mfcc+f_0	f_0	Mfcc+f_0	f_0	Mfcc+f_0	f_0	Mfcc+f_0	f_0
1	96.73	96.16	2.22	1.89	1.03	0.47	2.00	1.68	0.83	0.58
2	1.90	2.07	95.01	93.52	4.19	3.95	0.82	0.52	1.90	5.13
3	0.15	0.22	1.61	1.65	82.86	84.49	2.83	6.17	4.06	5.38
4	1.23	1.06	0.74	0.66	10.10	5.81	93.47	86.08	12.83	9.60
5	0.00	0.49	0.42	2.28	1.83	5.28	0.88	5.55	80.38	79.30

3. Tonal Variations & Contextual Factors

3.1. Role of spectral features

In Section 2.4, both spectral and f_0 features are used in CDTMs, although, f_0 features are normally considered as the main factors to distinguish tones. To investigate the role of spectral features in tone modeling, we train a new set of CDTMs with f_0 related features only. To get rid of errors caused by insertions and deletions in the first-pass decoding, two sets of models are compared on base syllables and boundaries generated from the transcription of the testing set at this time. The tone ERRs with/without MFCC are 7.9% and 10.7%, respectively. The tone ERR increases by about 3% after removing the spectral features.

The confusion matrixes of tones in the two experiments are shown in Table 2. We find that the spectral features are most helpful for recognizing T4. When only f_0 features are used, errors scatter into many cells at crosses between T3, T4 and T5. After spectral features are added, errors become concentrated in two cells, i.e. T3 and T5 are more likely to be

recognized as T4. When looking into the two types of errors, we further find that about half of them are at word final (SP4) or phrase final (SP5) positions. An intuitive explanation for this is that, at these positions, pitch range is reduced so that all the three tones exhibit a low-falling f_0 pattern. They are not easy to be distinguished by f_0 features. Besides, the speech energy at these positions is normally weak so that spectral features cannot provide enough discriminative information as well. Since there are much more T4 syllables than T3 or T5 syllables in the training data, CDTMs for T4 are better trained so that they tend to absorb T3 and T5 syllables that have f_0 patterns similar to T4.

3.2. Ranking contextual factors from decision trees

When we build CDTMs with multiple factors, each of which can take multiple values, the total number of possible contexts is huge. It is impossible to have a model for all of them. Therefore, decision-tree based clustering is performed to group states with similar context. The growing of trees, to a great extent, reflexes the impact of contextual factors. The question used at the root node normally causes larger variations than others and questions close to the root node are more important than those far away from the root node. Therefore, the importance of a question is weighted by:

$$W_{Qi} = 1/n, \quad (1)$$

where n represents that question Q_i is used for the n^{th} splitting of a tree.

The importance of a contextual factor is then calculated as the sum of weights of all questions regarding to it and is divided by the total number of values the factor can take. For example, for SP, the denominator is 6 and for LC and RC the denominator is 54.

The weights of factors on three states of five Mandarin tones are shown in Figure 2. Since the weights for LC and RC are all very small, they are not plotted.

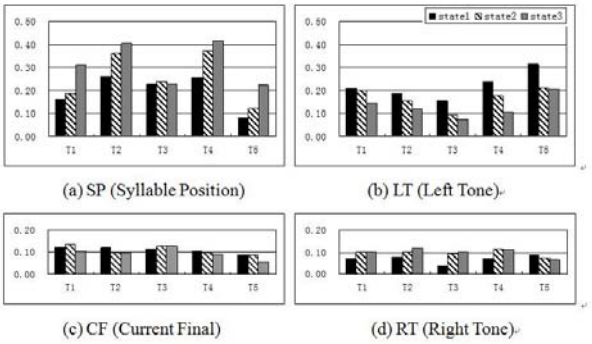

Figure 2. *The importance of contextual factors (by tones and by states)*

We find that, in most situations, SP has the largest weights and LT ranks the second. The impact of SP on three states of tones increases from left to right, except for T3, at the same time, the impact of LT decreases. In other words, the pitch level of the first state is normally determined jointly by SP and LT, while, the pitch level of the last state is mainly constrained by SP. CF has rather stable impact on the three states of T1-T4. RT has some influences on the second and third states of T1-T4. For T5, the neutral tone, LT is the most important factor and SP mainly influences its last state. These observations match previous studies on neutral tone. For T3, CF has larger impact than LT on the second and third states.

Since SP and LT are the most important factors that contribute to tonal variations, they are analyzed in details in the following sub-sections.

3.3. Tonal variation caused by syllable positions

The stylized f_0 patterns of five tones at six positions are shown in Figure 3. The upper bound of pitch range decreases when its position changes from phrase initial (SP1) toward phrase final (SP5). The lower bound is rather stable except for SP1, where it is much higher. T4 mainly occupies the top half of the pitch range at SP1, SP2 and SP3, while it moves toward the bottom half at SP4&5. The onset pitch of T2 is normally close to that of T3 and its offset does not reach as high as T1, i.e., in continuous speech, the high target of T2 is not easy to achieve. T5 normally locates closely to T3 except at SP3, where it is at the middle part of the pitch range. All tones at SP6 demonstrate some distinct features. For example, they have the smallest pitch range and both upper bound and lower bound locate at the middle part of the overall pitch range of the speaker. The relative position of tones at SP6 is the most similar to that of isolated syllables.

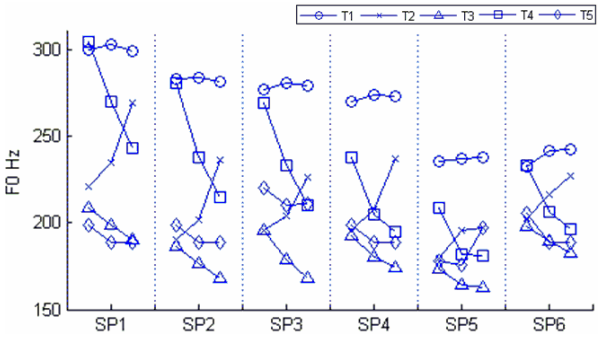

Figure 3.Stylized f_0 patterns of five tones at six positions

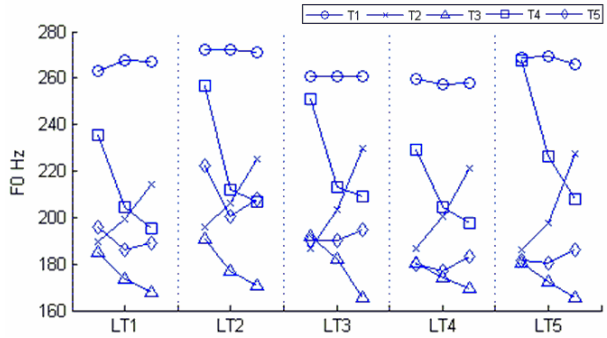

Figure 4. Stylized f_0 patterns of five tones preceded by different tones.

3.4. Tonal variation caused by left tones

The stylized f_0 patterns of five tones are plotted by their left tones in Figure 4. The pitch range of 5 tones is rather larger when they are preceded by a neutral tone. A main reason is that most T5 appears at word final so that tones after T5 are very close to tones at SP2. So, in this section, we exclude them from analyses.

T1 has high pitch when it is preceded by T1 or T2, which has a high offset, and its pitch level steps down when it follows a low offset tone (T3&T4). T4's onset is the highest when it follows T2 because the turning of f_0 from rising to falling takes additional effort and it is the lowest when preceded by another T4. An interesting observation is that the offset of T4 is higher when it is preceded by T3, a low offset tone, than when it is preceded by T1, a high offset tone. A possible reason is that T3 has reached a rather low pitch so that it tends to push up the onset of the falling tone. T2 has higher onset and lower offset when it is preceded by high offset tones (T1&T2). The onset of T3 is lower when it is

preceded by a low offset tone (T4&T5). T5's onset is the highest when it follows T2 and is the lowest when it follows T4&T5. After T3, T5 shows a flat pitch pattern, matching the findings in previous studies.

3.5. Joint impact of syllable position and left tone

From Figure 2, we find that onset pitch of tones is constrained jointly by SP and LT. The joint effect between the two factors is illustrated by simplified decision trees. Figure 5 shows the top five questions for splitting nodes in the decision trees of the first states in four tones. The five questions cluster all state-in-context into five groups and the context information and f_0 mean of each group are given in rectangular boxes. We can see that states-in-context are normally first clustered by their positions, especially at phrase initial/final or not. Within each sub group, states are further clustered according to its left tone. The high offset tones tend to have similar impact on the following tones except for T4. The onset pitch of T4 is normally lower when its left tone is T1 or T4 at SP4, SP5 and SP6. Yet, such a tendency is not observed at other positions.

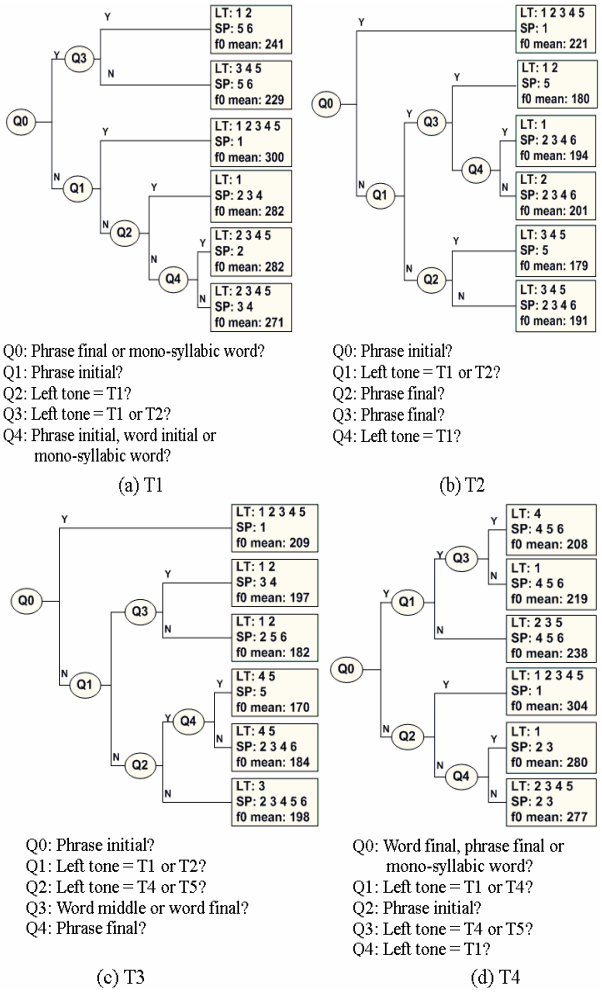

Figure 5. Joint effect of SP and LT in clustering the first state of four tones

4. Conclusions

In this paper, we train context-dependent HMM from a large speech corpus to represent tone-in-context. We haven't achieved any gain in tone recognition task, when CDTMs are compared with traditional tri-phone HMMs. However, CDTMs depict a clearer picture on how tone varies in different context. Therefore, it is a powerful tool for studying tonal variations.

Among the six contextual factors we explore, syllable position is the most important one and left tone ranks the second. Both right tone and current final influences the pitch level of tones, but their impact is weaker. The impact of left and right consonants is the weakest. We further find that the impact of a factor varies across tones as well as the three states of a tone. Normally, onset pitch of a tone is determined jointly by syllable position and left tone, while, the offset pitch is mainly depended on syllable position.

5. Acknowledgements

The authors would like to thank Yao Qian and Yu Shi for their helps on training CDTMs.

6. References

[1] WU, Z. J., "Tone-Sandhi in Standard Chinese Sentence", Zhongguo Yuwen, 1982, (6), 439-499. (In Chinese)

[2] Wu, Z. J., "Tone-Sandhi Patterns of Tri-syllabic Combinations in Standard Chinese", Zhongguo Yuyan Xuebao, 1984 (4), 70-92. (in Chinese)

[3] Xu, Y., "Contextual Tonal Variations in Mandarin", Journal of Phonetics, Vol. 25, 61-83, 1997.

[4] Shen, J., "Pitch Range of Tone and Intonation in Beijing Dialect", Experimental Phonetic Study of Beijing Dialect, Lin, T. et. al. edited, Peking University Press, 73-130, 1983. (In Chinese)

[5] Chu, M. and Lu, S. N., "A Text-to-Speech System with High Intelligibility and High Naturalness for Chinese", Chinese Journal of Acoustics, English Edition, Vol. 15, No. 1, 81-90, 1996.

[6] Wang, H. L., Qian, Y., Soong, F. K., Zhou, J. L. and Han. J. Q., "Improved Mandarin Speech Recognition by Lattice Rescoring with Enhanced Tone Models", In: Huo, Q., Ma, B., Chng, E.S. and Li, H.Z. (eds): Chinese Spoken Language Processing. Lecture Notes in Artificial Intelligence, Vol. 4274, (Proc. of ISCSLP-2006, Singapore), Springer-Verlag Berlin Heidelberg, 2006, 736-747.

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

[8] Chu, M., Peng, H., Yang, H. Y. and Chang, E., "Selecting Non-uniform Units from a Very Large Corpus for Concatenative Speech Synthesizer", Proc. of ICASSP, 2001.

[9] Tokuda, K., Masuko, T., Miyazaki, T. and Kobayashi, T., "Hidden Markov Models based on Multi-Space probability Distribution for Pitch Pattern Modeling", Proc. of ICASSP, 1999.

[10] Wang, H., Qian, Y., Soong, F. K., Zhou J. and Han J., "A Multi-Space Distribution Approach to Speech Recognition of Tonal Languages", Proc. of Interspeech, 2006.

[11] Zhou, J., Tian, Y., Shi, Y., Huang, C. and Chang, E., "Tone Articulation Modeling for Mandarin Spontaneous Speech Recognition", Proc. of ICASSP, 2004.