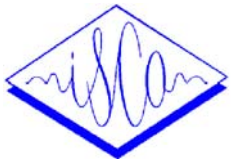

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

MODELING CARRYOVER AND ANTICIPATION EFFECTS FOR CHINESE TONE RECOGNITION

Jin-Song Zhang and Hiromichi Kawanami
Department of Information and Communication Engineering
School of Engineering, University of Tokyo
Bunkyo-ku, Tokyo, 113-8656, Japan
zjs@gavo.t.u-tokyo.ac.jp, kawanami@gavo.t.u-tokyo.ac.jp

ABSTRACT

This paper presents our new approach to model tone coarticulation of Chinese continuous speech for tone recognition. We suggest that coarticulation effects between two neighboring tones are rather unstable, since they may be uni-directional, bi-directional, or none despite of the same phonetic contexts. Instability is suggested due to non-local prosodic events like prosodic phrase boundaries or stress effects. Hence, we propose that context dependent tone models should be estimated according to the exact underlying coarticulation effects. To simplify label work for coarticulation effects, label sets as few as 3 labels are adopted. Also F0 contours of tone nuclei are used to facilitate human to discriminate tone coarticulation effects. A new search algorithm for the output candidates was also proposed to adopt the new modeling method of tone coarticulation effects. Preliminary experiments on a female's utterances of data corpus HKU96 showed the effectiveness of the new approach.

10.21437/Eurospeech.1999-181

1. INTRODUCTION

Lexical tone recognition is crucial for recognizing spoken Chinese and understanding its intonation. There are four basic lexical tones, each having a specific fundamental frequency (F0) contour. F0 contours of tones in isolated syllables are stable. However, contours in continuous speech vary due to coarticulation effects of human speech production. Since F0 contour is the key feature for tone recognition, these variations should be modeled to realize high performance.

Conventionally, context dependent (CD) tone models are used to model F0 contour variations due to coarticulation effects [1]. However, through experiments, we found recognition rate increase by the use of CD models is not so prominent for tone recognition as it did to phoneme recognition [2,3]. After comparing tone and phone coarticulation effects, we found there exist different coarticulation characteristics between them. The difference lies in that: coarticulation of phones within syllables or words is very stable [1], whereas that of tones is much more flexible. In other words, stability of tone coarticulation is more like that of interword phones than that of intraword phones. Therefore, introduction of CD models to tone recognizers did not bring a gain as high as it did to phoneme recognition.

Phonetic studies have revealed that there exist both directional coarticulation effects of tone production, i.e. forward carryover and backward anticipation effects [4,5,6]. And coarticulation exerts greater effects on tonal F0 contours when utterance has a faster speech rate [4]. Our observations over continuous utterances showed clearly that tone coarticulation effects are more flexible than those discussed in [4,5,6], since there they only tested trisyllabic sequences. This leaded us to reconsider the conventional modeling method of tone coarticulation effects. On one hand, CD tri-gram models assume the contemporary existences of carryover effect from preceding tone and anticipation effect from succeeding tone. On the other hand, underlying tone coarticulation effects of a tone in continuous speech may be uni-directional, i.e. either carryover or anticipation, or bi-directional, or no coarticulation. Therefore, in the conventional CD methods, a CD tone's acoustic model may be incorrectly estimated if the assumed coarticulation effects do not exist consistently in training data. Through our work on tone recognition of continuous utterances, we found this is true and is the possible reason leading to the low efficiency of CD tone models.

One plausible answer to the low efficiency of CD tone models is to train the CD models using training data with more accurately labeled tones considering their underlying coarticulation effects. Then another problem arises: how to label possible tone coarticulation effects in continuous speech. Considering the interaction of tone coarticulation and other prosodic events like stress and sentential intonation, one may find it too complicated to label all kinds of prosodic effects for continuous utterance. Instead, we suggested that presently we only need to label tone coarticulation effects between two neighboring tones, i.e. local effects, and left other non-local effects untouched. This is because not only labeling local tone coarticulation can be more reliable, but also tri-gram CD tone models only desire local effects from the preceding and succeeding tones.

The following sections present more details about our findings and work on tone recognition.

2. LEXICAL TONE RECOGNITION AND CONTEXT DEPENDENT TONE MODELS

Each syllable in Chinese corresponds to a morpheme and has basic structure of (C)V with a lexical tone. The four basic lexical tones (Tone 1, 2, 3 and 4) can be represented by their tone onset and offset F0 values, or their typical F0 contour shapes [Table 1]. Besides the four basic lexical tones, there is also a neutral tone which has no specific F0 contour.

	Onset F0	Offset F0	F0 contour shape
Tone 1	high (H)	high (H)	high and flat
Tone 2	low (L)	high (H)	rising
Tone 3	low (L)	low (L)	falling and rising
Tone 4	high (H)	low (L)	falling

Table 1. Characters of four basic lexical tones.

Acoustic features for tone recognition are usually F0, frame power and their 1st and 2nd time derivatives. Use of power features is because not only they really make effects, but also they showed to be strong cues for tone perception by human beings when F0 is missing [7]. Compared with power features, F0 features play more important roles in recognizing the lexical tones. F0 time derivatives can classify dynamic F0 contour shapes (rising, falling or flat). F0 itself is useful to differentiate Tone 1 from Tone 3, because either micro-prosody or tone coarticulation may result in a falling-and-rising shape for tone 1 or a flat shape for Tone 3.

Similar to segmental phone coarticulation, there is also coarticulation for tones since it also results from the physiological articulatory constraints. The inertial characteristic of a bio-mechanic system makes it impossible for its state to change abruptly from one to another quite different within a rather short period. As the result, surface F0 contours of connected tones may deform greatly from the canonical shapes of lexical tones in [Table 1]. Therefore, studying tone coarticulation is helpful for improving tone recognition performance.

Like modeling segmental phone coarticulation effects, it is intuitive to use context-dependent acoustic models to model tone coarticulation effects. For instance, a tone sequence "A– B – C – D" is acoustically modeled by context-independent (CI) or context-dependent (CD) models like:

CI:	A	B	C	D
CD:	A-(B)	(A)-B-(C)	(B)-C-(D)	(C)-D

Therefore, coarticulation effects on each tone due to its neighboring tones can be modeled by allotone models (CD models): bi-gram tone models used to model tones at the beginning and end of a utterance, and tri-gram tone models used to model tones in the middle of an utterance.

3. TONE COARTICULATION: UNIT-DIRECTIONAL OR BI-DIRECTIONAL?

One question associated with the CD models is that whether the assumed coarticulation effects always occur stably for the same phonetic contexts or not. In order to discuss this question, we adopt the concepts "Carryover" and "Anticipation" to describe tone coarticulation effects on the two neighboring tones.

- "Carryover" indicates the forward effect that happens when a tone's F0 contour is affected by its preceding tone.
- "Anticipation" indicates the backward effect that happens when a tone's F0 contour is influenced by its succeeding tone.

The two kinds of effects may occur either individually or simultaneously. Individual occurrences are uni-directional tone coarticulation effects; simultaneous occurrences are bi-directional. Although there is no general agreement on the directionality of Chinese lexical tone coarticulation effects, we say:

- Both uni-directional and bi-directional effects exist in Chinese lexical tone coarticulation despite of phonetic contexts.

We give evidences from two aspects:

1. From the disputations on whether tone coarticulation is "symmetric" or "asymmetric". In [5], Shen states clearly that tone coarticulation is bi-directional and symmetric: anticipatory and carryover are similar. However, Shen negated the existence of uni-directional coarticulation. In [6,7], Xu suggested that tone coarticulation effects are asymmetric in that carryover effects are mostly assimilatory and anticipation effects are mostly dissimilatory.
   - Assimilatory effect indicates that the offset of a tone and the onset of its succeeding tone will appear as if they have been assimilated or partially assimilated to each other.
   - Dissimilatory effect acts contrary to the assimilatory effect in that the offset of a tone and the onset of its succeeding tone appear as if they have departed from each other.

   Based on Xu's observations and analyses, one can deduce that if the carryover effect is assimilatory then the dissimilatory anticipatory effect cannot happen, and vice versa. Therefore there should be uni-directional coarticulation effects in Xu's data. Ignoring the disputations about the possible symmetric or asymmetric characteristics, one fact is clear that there are both uni-directional and bi-directional coarticulation effects observed in Shen's and Xue's experimental data.
2. More direct proofs are found from speech data used for our tone recognition experiments. Due to the differences between the trisyllabic sequences used in [5,6,7] and the real continuous utterances we use, there are more frequent appearances of uni-directional coarticulation effects in our data. For example, one of our findings about the uni-directional and bi-directional coarticulation effects is that they are strongly interacted with non-local prosodic events like prosodic phrase structure and utterance stresses. The findings can be summarized as:
   - Tone coarticulation effects are mostly bi-directional for tones belonging to a word or a word group which may be called as a prosodic phrase.
   - Uni-directional coarticulation effects often occur when there are some kinds of "break strengths" affecting the coarticulation of two neighboring tones. The "break strengths" can be possible phrase boundaries of various levels, or stresses on one of or both tones, and etc.
   - Strong "break strengths" may eliminate any

possible coarticulation effects, i.e. a tone being isolated from its neighboring tones in continuous speech.

Fig. 1 gives an example illustrating our findings about tone coarticulation in continuous speech. We can see surface F0 contours may be different, even the phonetic contexts are all the same, just because the tone coarticulation effect is bi-directional or uni-directional.

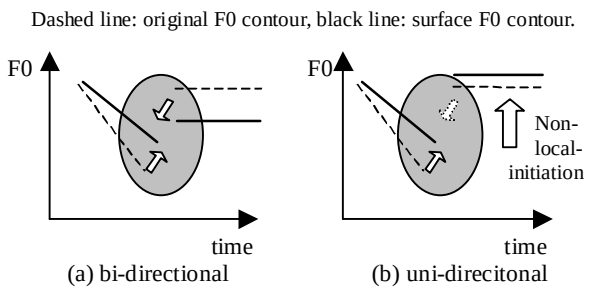

Fig. 1 Illustration of bi-direcational and uni-directional tone coarticulation effects through a pair of tones: Tone 4 –Tone 1. In (a), asimilatory effect acts bi-directionally so that the low offset of Tone 4 is raised to a little higher position than it originally should be, and high onse of Tone 1 is lowered. In (b) assimilatory effect on Tone 1 is stoped by some non-local initiation effect whereas preceding Tone 4 is still affected.

Now that we found there may be different underlying tone coarticulation effects for the same phonetic contexts, then we can say conventional ways adopting CD models for tones may incorrectly model the possible tone coarticulation effects in continuous speech. For example, in the tone sequence "A <-> B -> C <-> D", where "<->" indicates bi-directionality and "->" uni-directionality. By the conventional way, tones "B" and "C" are modeled as "(A)-B-(C)" and "(B)-C-(D)" respectively. For tone "C", both carryover and anticipation are correctly modeled. However, as tone "B" has no anticipation, the tri-gram "(A)-B-(C)" is incorrect. An improvement is to model tone "B" using a bi-gram "(A)-B". Therefore, a new scheme of CD model generation is proposed:

- Generation of CD model for a tone should only depends on the coarticulation effects it is really affected, but not on its position in the utterance.

4. LABELING UNI-DIRECTIONAL AND BI-DIRECTIONAL TONE COARTICULATION

A speech database labeled with accurate coarticulated tones is necessary to train tone models for recognition. Although number of possible kinds of tone labels exceeds 200 considering the neutral tone and phrase-final silence, they can be automatically generated if the four kinds of coarticulation effects are labeled, i.e.

1. bi-directional,
2. carryover only,
3. anticipation only,
4. no coarticulation.

We used four "break indices" to label the four kinds coarticulation effects: "b" indicates a full break and no coarticulation; "b-r" indicates anticipation that comes from right direction in the time axis; "b-l" indicates forward carryover effect that affect from left to right; and none label means a bi-directional coarticulation. So 3 kinds of labels were actually assigned during labeling work.

With these understandings about tone coarticulation in continuous speech and the simplified label set, one may still feel frustrated when facing too complicated surface F0 contours. To facilitate the labeling work, we developed a tone-nucleus detection system [3]. "Tone nucleus", which we once proposed as a robust means for tone recognition [3,8], spans from onset to offset of a tone. By viewing tone nuclei, we ignore variations of "onset and offset courses" in a tone F0 contour due to tone coarticulation effects, and focus on the possible coarticulation effects on tone nuclei. Then the coarticulation effects may easily be recognizable.

4.SEARCH ALGORITHM

After coarticulation effects are modeled as either bi-directional or uni-directional, search algorithm through candidate array also need to be modified from the conventional one during tone recognition. This is illustrated in Fig. 2 for conventional and modified search methods for two tones "A-B", which locate in the middle of a sentence.

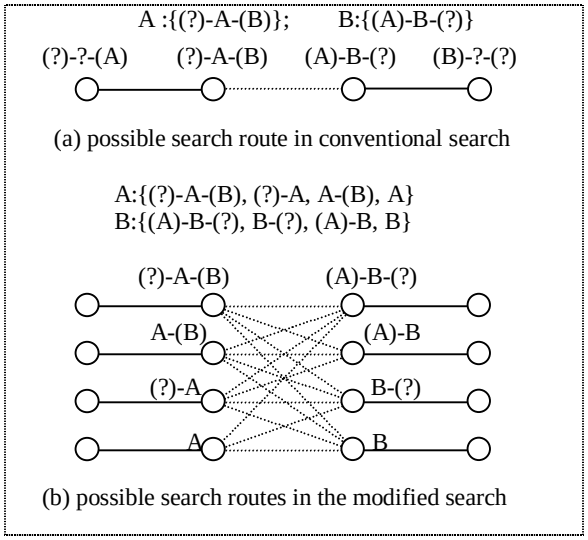

Fig. 2 Illustrations of conventional and modified search algorithms for the recognition of two tones "A-B" located in the middle of an utterance.

5.TONE RECOGNITION EXPERIMENTS

Tone recognition experiments have been carried out on a female speaker (0f) in data corpus HKU96. 500 utterances titled from cs0f0001 to cs0f0500 were used as training set, 200 utterances from cs0f0501 to cs0f700 were used as testing set. The number of syllables is 6419 for training set and 2567 for testing set. Tone coarticulation effects of the 500 training utterances were hand-labeled before tone recognition.

Continuous density HMMs are used as tone models. The number of HMMs in conventional CD method is 176 as in [1,3]. The number is 235 in the proposed "new

CD" method: 176 HMMs for the conventional method plus 59 additional HMMs for representing pause effects. And the number in context independent (CI) experiments is 6: 5 for lexical tones and 1 for silence. HMMs have left to right configuration. Number of states is 5 for the 4 basic tones, and 3 for the neutral tone and silence. Mixture number is 6 in the middle states and 2 in the beginning and ending states for the 4 basic tones, and less mixture numbers for the neutral tone and silence HMMs. Due to insufficient training problem, all the CD HMMs were first interpolated from CI HMMs and had tied transition matrices. Then they are re-estimated using training data with either conventional or newly adapted CD tone labels.

Tone recognition based on the following 3 systems have been carried out to examine the effect of the new CD method against the conventional CD method. The results of CI experiments act as the probable reference baselines for CD method evaluation.

- System A: using features of full syllable F0 contour.
- System B: using features of tone nuclei.
- System C: introducing a new modeling technique for "downstep" effects to system B.

Acoustic features used in systems A and B include F0, rms power, and their first and second time derivatives. For system C, besides the features used in systems A and B, additional two kinds of F0 features are used to model "downstep". Tone recognition accuracy for the testing set in all nine experiments are illustrated by Fig. 3.

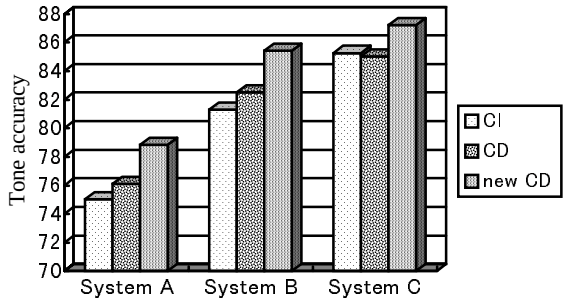

Fig. 3 Tone recognition experimental results. Where "CI", "CD" and "new CD" indicates context independent, context dependent and the proposed method.

Findings based on the experimental results include:

1. Accurate modeling of tone coarticulation effects has brought in the highest performances for tone recognition in all 3 systems. This indicates the ideas suggested in this paper do make positive effects on tone recognition.
2. The improvement of tone recognition performance also indicates the coarticulation effects are properly labeled.
3. The results of system C are very interesting. After we adopted a new modeling technique for "downstep" effect (we will discuss it in a future paper), conventional CD method brought no improvement compared with that of CI system. However the proposed new CD method still increases the performance. This can be explained as follows: Since the modeling technique for the specific coarticulation effect "downstep" is effective (in CI case, about 4% higher than the rate for system B, 10% higher than the rate for system A), further incorrect modeling of local coarticulation effects by CD models may lower its performance, whereas correct modeling by new CD models may still improve the performance.
4. Further improvements are still possible. They lie in that: (1) There exists a insufficient training problem, which can be solved when training set is enlarged. (2) From the detailed results, we found the average accuracy for four basic tones is already higher than 90%, whereas that of the neutral tone is rather low (around 30%). This is because both F0 and frame power do not serve as effective discriminating features for the neutral tone. The neutral tone should be dealt by other means. (3) Due to human factors, it is questionable whether the coarticulation effects are labeled consistently. We are searching a way to label them automatically or semi-automatically.

5. CONCLUSIONS

We have developed a new approach dealing with local tone coarticulations for tone recognition. We will test it on larger database and apply it to speaker independent task in the future. Also we want to point out, since the dealt coarticulation effects are due to physiological mechanism, it is reasonable to assume it applicable to other languages, especially tonal languages.

ACKNOWLEDGEMENTS

Special thanks to my thesis advisor, Professor Keikichi Hirose at the University of Tokyo, for his kind advice and encouragement during the work.

We also express our appreciation to Goh Kawai for his suggestions and proof-reading, Jinfu Ni for the discussions, and other colleagues at Hirose Lab of the University of Tokyo for their generous helps.

REFERENCES

[1] H.M. Wang, et al, "Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulaary using limited training data". IEEE Trans. on SAP, 5, No.2, 1997,195-200.

[2] L. Rabiner and B.H. Juang, "Fundamentals of speech recognition", Prentice Hall PTR, 1993, 458-476.

[3] K. Hirose and J.S. Zhang, "Tone recognition of Chinese continuous speech using tone-critical segments", Proc. of EuroSpeech99, Budapest, Hungary, Sept. 1999.

[4] X. N. Shen, "Tonal coarticulation in Mandarin", Journal of Phonetics 18, 1990, 281-295.

[5] Y. Xu, "Production and perception of coarticulated tones", J.A.S.A. 95 (4), April, 1994, 2240-2253.

[6] Y. Xu, "Contextual tonal variations in Mandarin", Journal of Phonetics 25, 1997, 61-83.

[7] D. H. Whalen, and Y. Xu, "Information for Mandarin tones in the amplitude contour and in brief segments", Phonetica 49, 1992, 25-47.

[8] J.S. Zhang and K. Hirsoe, "A robust tone recognition method of Chinese based on sub-syllabic F0 contours", ICSLP98, Sydney, Australia, Dec. 1998,703-706.