MODELING DURATION AND TONAL COARTICULATION IN A MANDARIN CHINESESE SYNTHESIS

Hongwei Ding and Joerg Helbig
Technical Acoustics Laboratory
Dresden University of Technology, 01062 Dresden, Germany
FAX: +49 351 4637091, E-mail: helbig@eakss1.et.tu-dresden.de

ABSTRACT

We present in this paper the results of a duration study and a tonal coarticulation study designed for the concatenative Mandarin Chinese synthesis system developed at the Dresden University of Technology. It is reported that the duration model and the tonal coarticulation model are the two most important components of the prosody control in Mandarin. The material for the study of the two prosody components was extracted from a phonetically and prosodically labeled sentence database uttered by the speaker of the synthesis inventory. This approach ensures a high adaptation of the prosody control algorithm to the speaker-dependent characteristics of the synthesizer's voice, and turned out to be essential for the improvement of naturalness and user acceptance of the synthesized speech.

1. INTRODUCTION

A correct prosodic modeling plays an important role in intelligibility and naturalness in the synthesized speech, especially for tonal languages such as Chinese. There exist many references from publications on these items ([1], [2], [3], [4], [6]). Each of them has provided us with some useful references. But with a detailed research in these literatures, we have found:

- The duration studies have been conducted by various institutions ([1], [2], [6]), but the reported results of duration for every single phoneme are quite different, which may be due to different reading conditions.

- The tonal coarticulation studies have also achieved much attention in Mandarin synthesis in the past years ([3], [4]). But most of these enlightening results were expressed in verbal descriptions of a set of rules and they have to be completed by parametric knowledge for the prosody algorithms of the synthesis system.

Because of the special design of our synthesis inventory [5], which is coarticulation-motivated syllables with natural tone contours, the statistical results from these available information can not be directly employed by our system. An investigation for our own synthesis seems necessary.

Besides, the majority of these investigations have been carried out on the basis of isolated words or word chains. Our experiences in this area suggest that the results from the fluently spoken natural speech are of more importance for the design of prosody models of a synthesis system. There are, however, several publications, such as [2] for duration and [4] for tonal concatenation, which have reported on the successful investigation with naturally spoken speech database. These studies have shown us the possibility to study the effects of duration and tonal coarticulation from the material of the natural speech.

2. SYNTHESIS SYSTEM

The speech synthesizer is designed as a concatenative system on the basis of syllables. The speech inventory consists of 2910 naturally spoken syllables extracted from carrier sentences of a male speaker with standard pronunciation. It contains all Chinese syllables of about 1218, including their tones. Additionally, the vowel ending syllables were recorded in 3 coarticulatory environments (labial, alveolar and velar) in order to enable a correct modeling of the cross syllable coarticulation effects [5].

The prosody of the syllables fetched from the inventory will be modified in the time domain according to various models of duration, tonal coarticulation and sentence intonation. The study of the first two models is the purpose of the paper. The third model of sentence intonation combines the suprasegmental intonation pattern of the phrase and the segmental tonal structure of the syllables. The phrase intonation pattern is calculated by an algorithm using a linear declination model with accentuation facilities and a final fall/rise modification depending on the mode of phrase. The segmental intonation contours of the 4 tones are unit inherent in the synthesis inventory. They are shifted to the phrase intonation line at tone-dependent synchronization's points, so as to maintain their natural tonal contours. Thus the resulting pattern can be regarded as a superposition of the two intonation components [6].

3. MATERIAL FOR STUDY AND METHOD FOR ANALYSIS

Two kinds of material were chosen for the analysis of duration and tonal coarticulation

- The labeled inventory of the synthesis system
  2910 syllables with 5532 sounds

These syllables were segmented from carrier sentences [5], in which all the target syllables were followed by neutral tones. In such a way, all the syllables keep the constant prosody and their tonal contours are least influenced from the following tones. These characteristics proved to be very important for an inventory for concatenative systems in time domain.

- The naturally read texts
  61 sentences with 15 to 86 syllables, altogether 3054 syllables with 4962 sounds in 1756 words

The natural speech material of the text database was selected from newspaper articles. The style of the material is mostly news reports. The speaker was asked to read these articles smoothly and not overarticulated at a normal, fluent speaking rate. The sentences were selected from the recording of one hour, in order to get a statistical balance and a good coverage of the various parameters which can have influences on the duration of the phonemes. For the sake of tonal coarticulation study, besides the considerations mentioned above, the balance of the possible combinations of the lexical tones was also taken into account. The selected sentences contain 511 disyllabic words and 256 trisyllabic words, which were analyzed for tonal coarticulation effects. The sentences were manually labeled by a native speaker.

The labeling includes phonetic labels (phoneme symbols with their lexical tones) and prosodic labels. The prosodic labels include the prominence levels (3 prominence levels for the syllables, that is the sentence-, phrase- and word- stress); position of the phoneme (the boundary indication, that is the syllable-, word-, phrase- and sentence- boundary) and together with the mode of phrase (declaration, question etc.). Moreover, the F0-contour was calculated synchronously to the waveform of the sentences. For the duration study, the labeling information together with the derived phoneme duration values were statistically analyzed. For the tonal coarticulation studies, the measurements were obtained from graphical plots of the waveform and pitch tracks with an interactive interpolated sentence intonation contour. This gave us the possibility for a separate treatment of suprasegmental and segmental intonation components according to our prosody model.

4. MODELS AND RESULTS

4.1 Duration studies

The duration study was conducted with the above-mentioned material. The main purpose is to find :

- What is the relationship of our result to those reported in the literature? And
- What kind of factors have influences on the phoneme duration and to which degree should these factors be considered?

4.1.1 Duration Model

As a duration model we choose an approach similar to that adopted in [2], which is based on a multiplicative combination of the inherent phoneme duration values and contextual influence factors. In addition to the model proposed, we introduced a correction factor ($KORR_{(ST,PH)}$) for syllable durations. This factor is dependent on the type of phoneme (PH) and the type of current syllable (ST), which serves to avoid too large time scale modifications and thus ensures a continuous rhythm. The duration model [8] is represented by

$$DUR_{(PH)} = KORR_{(ST, PH)} \times (IDUR_{(PH)} \times F2 \times ... \times F13)$$

The structure of the inventory and the natural speech database makes it possible for us to analyze the following factors of influences, which are reported important in [2] on phoneme durations:

- (F1, F2) the actual phoneme and its lexical tone
- (F3, F4) phoneme type and tone of the preceding phoneme
- (F5, F6) phoneme type and tone of the following phoneme
- (F7, F8) number of the preceding and following syllables in the word
- (F9, F10) number of the preceding and following syllables in the phrase
- (F11, F12) number of the preceding and following syllables in the sentence
- (F13) accentuation (word-stress, phrase stress, sentences-stress, no stress)

The value of influence factors (F1 ... F13) were calculated according to the statistical analysis from the natural database. They were prepared for the calculation of the phoneme duration ($DUR_{(PH)}$) on the basis of the inherent duration of the phoneme ($IDUR_{(PH)}$) from the inventory of synthesis.

During the synthesis process, all these factors except for the accentuation can be obtained from the text analysis. The word stress can be obtained according to certain rules, but the phrase stress and sentence stress are still manipulated by hand. The information obtained from the text analysis will help to find the appropriate value of the factors of influence, and the above mentioned multiplicative model will then calculate the duration value for the concerned phoneme.

4.1.2 Results of duration study

In order to present the results from the analysis, the following several diagrams were selected for you to catch a general view over the investigation.

4.1.2.1 Comparison of duration values

In Diagram 1, one can find the mean duration values of several phonemes from different sources. These phonemes are of different types, ranging from „unaspirated plosive“, „unaspirated fricative“ to „fricative“ and „vowel“. The available sources are the reported duration values from AT&T [2], Ac.Taiwan [1], and the results from our inventory (INV) and our natural speech text database (TXT).

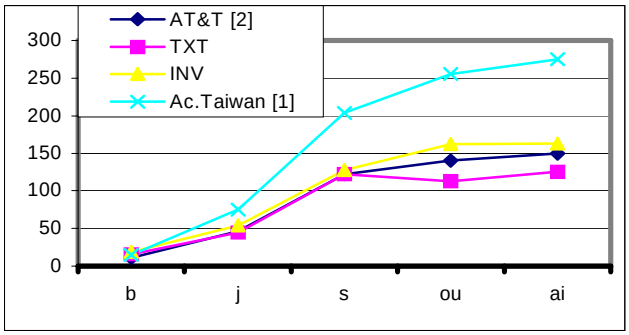

Diagram 1: Mean duration values in ms of several selected phonemes from different sources

It can be found that the results from different sources show the same tendency of the phoneme duration, the difference on the absolute value may be resulted from different speakers and reading conditions. The comparison of these phonemes reveals that the duration values of the phonemes from our inventory are situated between the reported results from the other two literatures. This convinces us that the duration value from our inventory can be used as the reference for the duration manipulation.

Whether the inherent duration values from the inventory can be very well manipulated by the statistics from the natural speech popped up as the next consideration. A comparison of the duration values of the phonemes from the inventory and those from natural text database is of great interest to us. This is illustrated in Diagram 2.

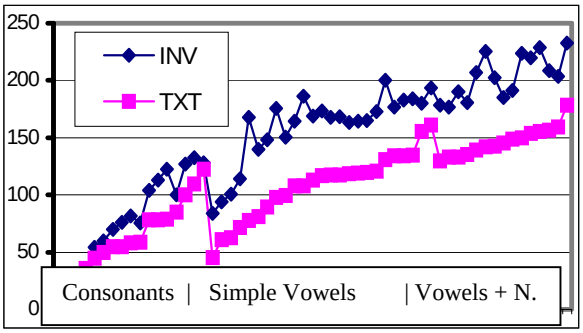

Diagram 2: Duration value in ms of all the types of phonemes (consonants, vowels and vowels with nasal ending) from the inventory of synthesis (INV) and those from natural speech text (TXT)

Although the reading conditions were quite different in the inventory from that in the text, the results keep the same order of the duration for all these phonemes. Except that the duration from the text is proportionally shortened as expected. This will further give us good reasons to generate a duration model from the statistics of natural speech on the basis of the inventory.

4.1.2.2 Duration for vowels

The factor that has the greatest influence on vowel duration is the prominence of the phoneme. This is clearly illustrated in Diagram 3.

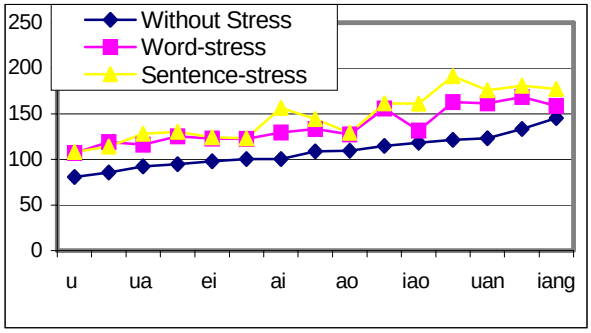

Diagram 3: Duration value in ms of vowels with different levels of stress from text (not all the vowels)

The accentuation of the syllable will largely enlarge the duration of the concerned phoneme. This suggests that in the process of synthesis, the stress should be clearly labeled for the syllable: such as sentence stress, phrase stress or word stress. Because of the small database of the natural speech and the fact that the sentence stress always tends to fall on certain vowels, it is not possible for every vowel to have the chance to carry the sentence stress. In the diagram only several vowels can be presented to show the influence of different levels of prominence on the duration of the vowels. But the fact that the vowel with stress is much longer than the same vowel without stress can be clearly demonstrated for almost all the vowels in the statistics.

Many such diagrams of comparison can be obtained to explain the influence of various factors on the vowel duration. To present the results concisely, the following verbal expressions can be concluded:

- Vowel together with nasal ending > Vowel without nasal ending
- Single vowel in the syllable > Vowel in a syllable with nasal ending
- Vowel at word final > Vowel at word initial
- Vowel at phrase end > Vowel at non-phrase final

(“>” means “is longer than”)

But the duration of the vowels from the natural speech does not show clear relationship to the tones of syllable, although it is clear that the third tone of the same phoneme is always the longest from the isolated syllables.

4.1.2.3 Duration for consonants and closure pauses

The results of the duration for the consonants show the agreement to the results of the other reports ([1], [6]), that the duration of the consonants is closely related to the manner of articulation. Besides, it is found that the consonant at the word final is longer than that in the word middle. What of great interest to us is that the closure of the consonant at the word initial is also longer than that in the word middle. This is shown in Diagram 4. This result can serve as the information for word segmentation.

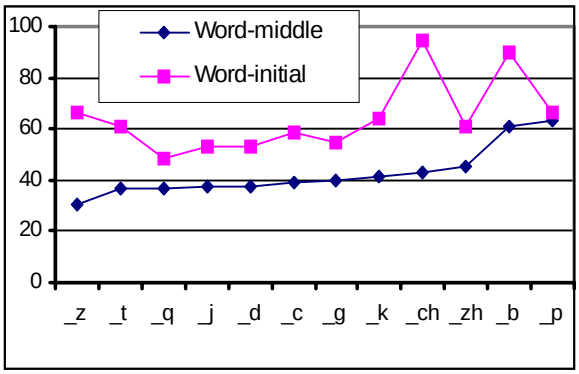

Diagram 4: Duration value in ms of the consonant closure pause from the text.

The duration study were analyzed with great care, and the statistical results were satisfactory.

4.2 Tonal coarticulation studies

When the syllables are connected together in natural speech, two kinds of tonal changes will take place

- The total change into another tone (sandhi rules)
- The modification of the tone contour without qualitative change (tonal coarticulation)

The first case of sandhi rules is realized by choosing another correspondent second tone, but the second case is much more complicated. Our attention will be focused on the second kind, which is responsible for most of the tonal changes when the syllables concatenated one to the other. The purpose is to investigate the effects of tonal coarticulation in natural speech, and generate rules according to the statistical results for the sake of synthesis.

4.2.1 Tonal coarticulation model

In order to generate a right intonation for the sentences, the suprasegmental sentence intonation declination line should be superimposed with the natural tonal contours of the syllables from the inventory. This is realized by the model introduced in [7], in which a synchronization's point is found in each of the four tones.

But such manipulation is far from enough to ensure the continuity of the intonation of the syllables in a word. It is to say that the tonal concatenation model is of much more importance than the sentence intonation model to guarantee the smooth processing of the intonation. In order to formulate the tonal coarticulation model from the statistics, the following parameters will be needed for the analysis with the assumption of a nearly correct separation of the suprasegmental intonation contour from the segmental tone contour [9]:

- The shift of the whole contour F0-DELTA (Distance of the F0-Synchronization's point of the syllable from the declination line of the sentence intonation)

$$F0_{DELTA} = (F0_{SYNC} - F0) / F0_{SYNC} * 100\% \quad \text{for } t = T_{SYNC}$$

T_{SYNC}: For Tone 1 and Tone 3 the beginning of the syllable
For Tone 2 and Tone 4 the time in the middle of F0 between MAX and MIN

With

$$F0_{SYNC} = F0_{PHRASE} + 0.1 * F0_{PHRASE} \quad \text{for Tone 1}$$
$$= F0_{PHRASE} \quad \text{for Tone 2 – 4}$$

- The range of fundamental frequency of the syllable (the modification of the F0 range of the syllable)

$$F0_{RANGE} = (F0_{SYL\_MAX} - F0_{SYL\_MIN}) / ((F0_{SYL\_MAX} + F0_{SYL\_MIN}) / 2) * 100\%$$

These parameters will be obtained from the analysis of the syllables in polysyllabic words. They depend on the sequence of lexical tones. That's why the polysyllabic words were separated into different groups. During synthesis, the natural tone contours of the syllables are modified by the statistics of these parameters.

4.2.2 Results of tonal coarticulation

The syllables were first divided into groups of bisyllabic or trisyllabic words. The tonal concatenation analysis was focused on the bisyllabic and trisyllabic words, for such combinations occupy a large part in Chinese word formation. The following will only show the results from that of the bisyllabic words, which is representative in Chinese. Except for the combination with neutral tones, which were separately studied, the combination of bisyllabic words fall into 15 groups.

First of all, the position of the first syllable in relation to the sentence declination line was calculated. This is presented in Diagram 5.

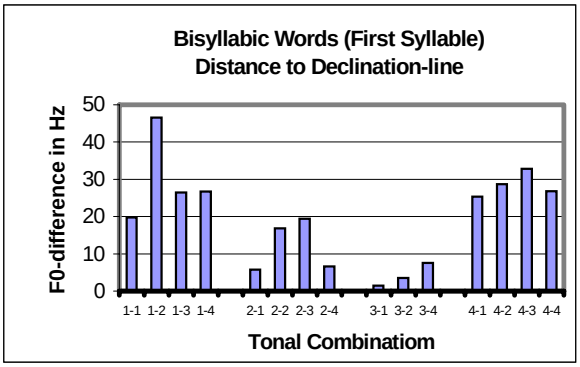

Diagram 5: The distance of the synchronization's point of the first syllable to the declination line.

The positive value explains that the first syllable is always higher than the declination line.

In order to decide the position of the second syllable, the distance of the synchronization's point of the syllable to the declination line and the difference of F0-MAX of the second syllable to the F0-MAX of the first syllable were also calculated. Diagram 6 is devoted to illustrate these two kinds of values.

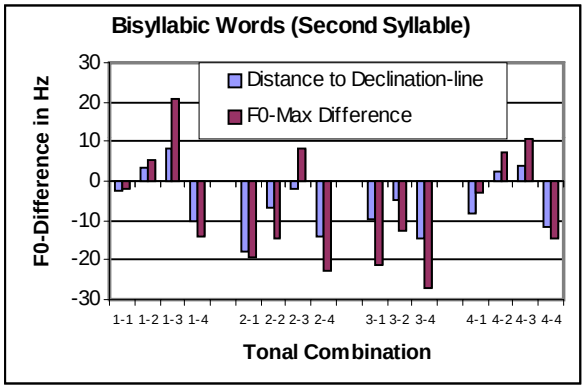

Diagram 6: The distance of the synchronization's point of the second syllable from the declination line, and F0-difference of the second syllable to F0-MAX of the first syllable.

Changes in tonal coarticulation were not only expressed by lowering or raising the whole tonal contour of the syllable, but also by enlarging or shortening the range of the fundamental frequency. To further decide the shape of the tonal contours of the first and second syllable in bisyllabic words, the F0-range of the syllables were also depicted in Diagram 7.

Owing to the complex factors interacting the effects, the analysis of the tonal coarticulation was not so carefully sorted as that in the duration study. For example the material was roughly analyzed without paying attention to the prominence of the syllable, which resulted in the great differences of the F0-Difference and F0-Range. This will be improved in the future research.

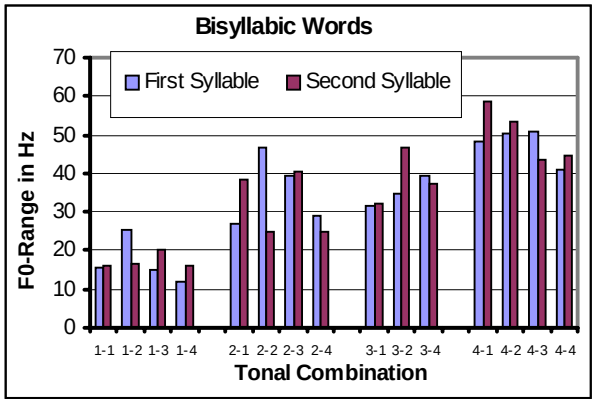

Diagram 7: The F0-Range of the first syllable and that of second syllable.

With the help of the results from these diagrams, the parameters of the above mentioned formulas can be calculated. The tonal coarticulation model works very well together with the model of sentence intonation, which has made great improvement on the prosody of the synthesis system.

5. CONCLUSION

On the basis of our studies concerning the duration and the tonal coarticulation effects of Chinese language, we have developed knowledge and experiences for the prosody algorithms of our synthesis system. A listening test showed an improvement in the naturalness and the fluency of the synthesized speech.

Future work will be carried out to enlarge our natural text database in order to get a better statistical coverage and balance of the material. The tonal coarticulation effect across word boundaries also remains as a problem to be solved in the future.

6. REFERENCES

[1] Tseng, C. (1995): "A Phonetically Oriented Speech Database for Mandarin Chinese". In ICPhS 95 Stockholm. 326-329. 1995

[2] Shih, C. & Ao, B. (1995): "Duration Studies for the Bell Laboratories Mandarin Text-to-Speech System" In J. Van Santen etc. (eds.) Progress in Speech Synthesis, Springer 1995, p.383-399

[3] Wang, R. et. (1996): "A New Chinese Text-to-Speech System with High Naturalness". Proc. ICSLP 96. P.1441-1444

[4] Lee, L. et. (1989): "The Synthesis Rules in a Chinese Text-to-Speech System". IEEE Trans. Acous. Spe. & Sig. Proc. 1989. P.1309-1320

[5] Helbig, J. & Ding, H. (1997): "A Syllable-based Mandarin Chinese Speech Synthesis regarding Cross-syllable Coarticulation Effects". ICSP 97, Aug. 26-28.1997. Seoul, Korea. 173-176

[6] Qi, S & Zhang, J. (1982): "A Study of Duration of

Chinese Consonants". In Shengxue Xuebao (Acta Acoustica) Vol. 7:1. 8-13
[7] Ding, H. & Helbig, J. (1997): "Natural Tone Contours in a Mandarin Chinese Speech Synthesizer". Proc. ESCA Workshop on Intonation, Athens 1997. 95-98
[8] Ding, H. & Helbig, J. (1998): "Untersuchungen zur Dauersteuerung für ein chinesisches Sprachsynthesesystem" DAGA 98 (to be published)
[9] Ding, H. & Helbig, J. (1998): "Untersuchungen zur Tonkoartikulation für ein chinesisches Sprachsynthesesystem" 9.Konferenz Elektronische Sprachsignalverarbeitung, 123-126 Aug.31- Sep.1. 1998