

# DURATION MODELING FOR CHINESE SYSTHESIS FROM C-TOBI LABELED CORPUS

*Zhu Weibin, Shen Liqin, and Niu Xiaochuan*

Speech Group, IBM China Research Lab.Beijing, 100085, China  
EMail: {zhuweib, shenlq, niuxc}@cn.ibm.com

## ABSTACT

A set of labeling criteria, C-ToBI (Chinese Tone and Break Index) was redefined to annotate the prosodic event in continuous speech in a hierarchical structure. There're 4 layers, i.e., intonational phrase, intermediate phrase, word, and syllable layer. The prosodic structure and break index and stress index tiers represent the core prosodic events of an utterance. The stress index represents the degree of accent of the constituents in each layer. The break tier represents the degree of the juncture of each pair of constituents in each layer. A duration model was built from a reading style corpus labeled with C-ToBI. The factors affecting the duration of a given segment come from two relatively independent levels. First, in segment level, the phoneme of the segment and the context do influence the duration. Second, in super-segment level, the influences come from multi-layers, which include the location and the degree of stress and break in different layers. Those factors with the property of directional invariance form the feature vector that was as the input of the linear duration model. And the model was part of a synthesis speech system, and its parameters were estimated by the statistic approach.

## 1. INTRODUCTION

In spoken speech, there're so many factors that correlate with the duration of a given phone. These factors come from many level including the physiological/psychological level, and lingual level, and phonetic level. And we focus our research efforts on the factors in phonetic level here. The phone identity and its context situation in the word/phrase/utterance do affect the variance of the duration. As what had been disclosed in the former research, many of those factors have the property of directional invariance [1]. To make the speech generated by a TTS system to sound more natural, the effects of these factors should be quantified, i.e., it's needed a more accurate duration predictor to compute timing of segments from a given symbolic sequence, which could be built from the prosodic labeled database [2].

ToBI transcription system has been used to research on the English prosodic events for several years [3], and it's modified and used for other languages such as Japanese and Chinese also. Two ToBI systems for Chinese prosodic transcription had been published before. In [4], a break index tier with scale 0 to 5 is defined to represent the prosodic structure. The consistency between two transcribers was not high enough. And the tonal representation was unavailable yet. In [5] system, both of a tone tier and a stress tier are used to represent the intonation, each tonal pattern will be annotated in the tone tier. As Chinese

is tonal language, each syllable has its tonal pattern, it's not an economic approach to describe the intonation by every tonal pattern being marked. In this paper, we refine a C-ToBI system to describe the prosodic events in a hierarchical structure. In our system, an intonation phrase is decomposed into three layers stepwise, i.e., intermediate phrase, word, and syllable. In each layer, the break indices represent the different degree of continuity between the every pair of constituents in the layer, while the stress indices mark the emphasis degree of each constituent.

Our main goal is to develop a duration module as part of a TTS system. It's obvious that not all of the annotations (some of them were labeled manually in the training database) can be output and/or correctly predicted by NLP (Natural Language Processing) component in current TTS system. The problem can split into two aspects. First is due to the limitation of the units coverage in the training database. It can be solved by fixing the application domain while the samples collected in the database are enough for the special domain, and unit-classing based on the empirical assumption can handling "new" input never seen before also. The second is because of the performance of current NLP. For Chinese system, except an accurate prediction of prosodic markers in phrase level, which is essential for the natural synthesis speech and also the most far future goal, the more practical goal includes, a). a more accurate parser outputs the correct word boundary and perfect spelling sequence; b). a lexicon with accent marked; c). prosodic rules/models for some canonical sentential forms which are used frequently. In this paper, however, our attentions are focused on to build a model computing the duration of a given segment with the given prosodic symbols.

## 2. PROSODIC LABELING SYSTEM: C-TOBI

### 2.1. Principle of C-ToBI

We're guided by three principles to refine the C-ToBI. a). Flexibility, it should be able to capture the categories of prosodic phenomena even from a large corpus. b). Reliability, there's high agreement among different transcribers. c). Compatibility, it should be in machine-readable representation format, for sharing the transcriptions across sites and across hardware/software platforms, and for further applications in TTS and SAR.

In our C-ToBI labeling system, there're 6 tiers:

- (1). a prosodic structure tier

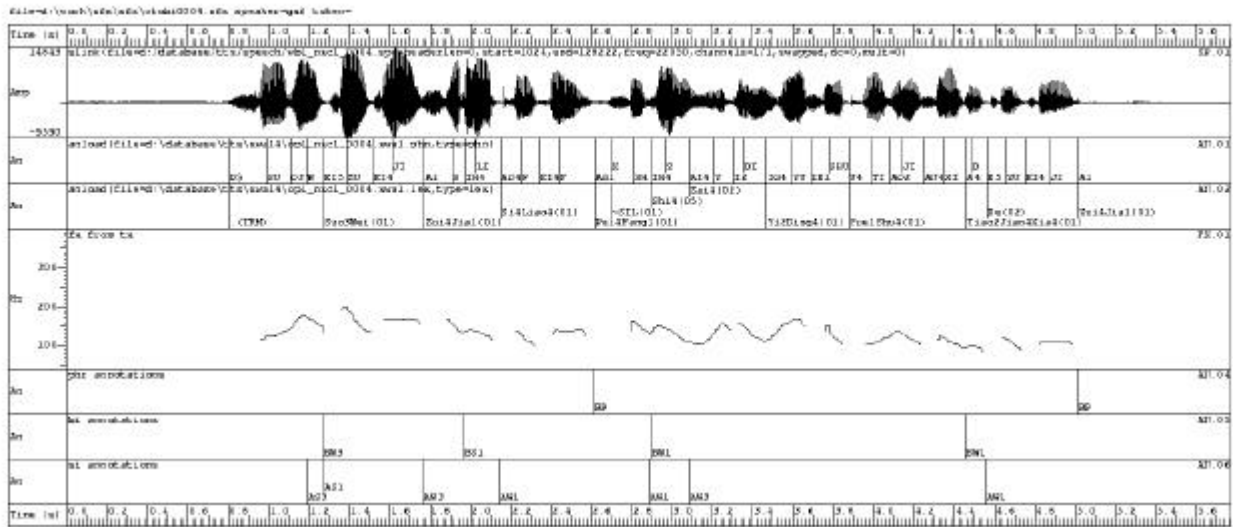


Figure 1. An example of C-ToBI transcription, scale 2 is not marked here

- (2). a stress index tier
- (3). a break index tier
- (4). an orthographic tier
- (5). a mood tier
- (6). a miscellaneous tier

The orthographic tier is a straightforward transcription of all of the syllables into ‘PinYin’ format in the utterance. The transcriber can use it as a reference, and the polyphone problem should be resolved in the tier.

The mood tier marks the utterance’s mood. Those mood patterns are classed into four types: indicative, imperative, interrogative, interjectional, by the pragmatics of the utterance.

The miscellaneous tier is used for any comments or markings, e.g., disfluency, laughter, audible breaths, and so on, desired by transcriber.

The prosodic structure and break index and stress index tiers represent the core prosodic analysis of Chinese. In Chinese, as a tonal language, different from intonation language, each syllable is with its tone, and the pitch contour of the utterance is the combination of the tonal pattern of each syllable and the mood pattern of whole sentence. The tonal pattern is modified by the different contexts from each layer, e.g. the break index between syllables in the word, the stress index of the syllable, the word, and the phrase, etc. Also the syllable is the basic unit of rhythm of spoken Chinese speech. So, the minimal prosodic unit in C-ToBI is the syllable.

Actually the prosodic structure based on syllable is relatively fixed for each word. In a word, the break index between the syllables and the stress index of the syllable normally is

invariant. Due to no authoritative lexicon with accent annotation marked, the work needs to be done manually now.

In most NLP (Nature Language Processing) system of Chinese, both for TTS or ASR system, the processing is based on the word. The specific situation in Chinese is that there is no word boundary in the normal Chinese character string. Sometimes it’s ambiguous to decide where a word boundary should be, especially by automatic method. It’s not seldom that the transcriber needs to modify the word boundary generated by Chinese parser. The accent of the word and the break index between the words will affect the pitch pattern of the word and the word duration as well.

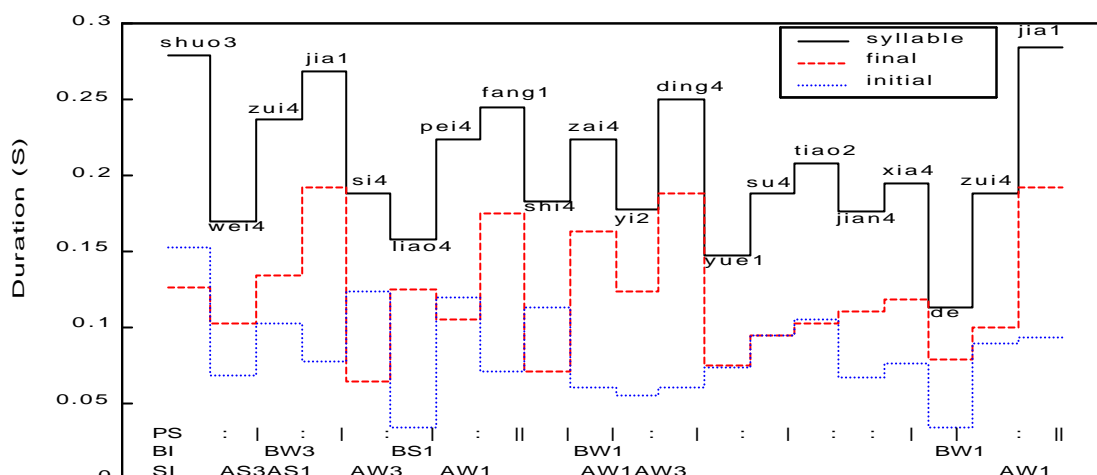
The words prosodic group consists the intermediate phrase, and both the break indices between phrases and the stress indices of the phrase do affect the intonation pattern of whole sentence – the intonational phrase.

Therefore, a hierarchical approach is chosen to represent the prosodic structure of Chinese spoken speech. There’re 4 layers: intonation phrase, intermediate phrase, word, and syllable. The different break/stress indices transcribe the continuity between/on the prosodic units of each layer.

The hierarchical structure used here, is simplifying the judgment of transcriber and keeping the compatibility with the speech system.

## 2.2 The Transcription of Core Prosodic Events

C-ToBI transcription for an utterance consists minimally of a recording of the speech, and an associated record of the fundamental frequency contour, and symbolic labels for prosodic events on the following 3 tiers.



**Figure 2.** Segment duration of one utterance labeled with C-ToBI, where “:”, “|”, “||” represents the syllable, word, and intermediate phrase boundary respectively, and “PS, BI, SI” represents the prosodic structure, break index, and stress index tier respectively.

(1). The prosodic structure tier

The utterance is transcribed in a hierarchy, from intonation phrase, down to intermediate phrase, then word, and then syllable. The boundary of syllable is obvious for Chinese. And the word boundary comes from the lexicon. The intermediate phrase marked by a words group with a perceived boundary which may be caused by sudden changes in pitch, duration, energy and pause.

(2). The break index tier

Break indices represent the degree of continuity perceived between each pair of syllables in a word, or words in a intermediate phrase, or intermediate phrase in a utterance. The scale is from 0 to 4.

- Scale 2 is the default value, for the normal break level perceived.
- Scale 1 is for reduced boundary, while scale 3 is for more prolonged boundary then normal.
- Scale 0 and 4 are for extremely situation, scale 0 for extremely reduced boundary and scale 4 for extremely prolonged boundary, both of scale 0 and 4 are seldom used for reading style speech.
- BP1, BW3, and BS2 represents the intermediate phrase boundary with scale 1, the word boundary with scale 3, and the syllable boundary with scale 2, respectively.

(3). The stress index tier

Stress indices represent the degree of stress perceived on a syllable in a word, or a word in a intermediate phrase, or a intermediate phrase in a utterance, and the pitch contour can provided the associative objective criteria. Also the scale is from 0 to 4.

- Scale 2 is the default value for normal stress utterance, which is perceived on each layer.

- Scale 1 is for neutral utterance, while scale 3 is for more stress utterance then normal
- Scale 0 and 4 are for extremely situation, scale 0 for extremely neutral utterance and scale 4 for extremely stress utterance.
- AP2, AW1, AS3 represents the intermediate phrase stress with scale 2, the word stress with scale 1, the syllable stress with scale 3, respectively.

A software package SFS (Speech File System, from UCL) was used as the UI for C-ToBI transcribing. Speech signal and fundamental frequency contour were shown to the transcriber, and associated with the phone and word boundary segmented by IBM ASR system. Those boundaries need to be modified manually sometime.

To do labeling, first step is to label the intermediate phrase boundary, and the transcription on the prosodic structure tier is closed then; second step is to mark the break indices in different layers; the last step is to mark the stress indices in different layers. Figure 1. gives an example of C-ToBI annotation.

### 3. DURATION MODELING

#### 3.1. Analysis of Segmental Duration

There're so many factors contribute to the duration of a given segment. Some of them are from the physiological and/or psychological level, they are speaker dependent features. The utterance's prosodic structure is affected by speaker's emotion and its variance is constrained by the speaker's articulator. Sometimes, although the script is the same, it's uttered in different styles corresponding with different semantics in lingual level, the intonations are different from each other

obviously. And on what we should focus are the events in phonetic level. Generally phonetic exploration can be divided into two relatively independent parts: supra-segment analysis, and segment analysis. In segment level, phoneme of the segment and its context do affect the duration. For initial, its manner, articulator affects the phone's duration. For final, its tone and phoneme (i.e., monophthong, diphthong, nasal coda) affects the duration. And there's an interactive influence between the initial and the final in a syllable.

In supra-segment level, the factors come from multi-layer (word, phrase), which include the location and stress indices in different layers, the break indices of different boundary. C-ToBI labeled database will be used to construct the duration model. Figure 2. represents the segment duration with C-ToBI annotation. From the figure, it can be seen that: break and stress indices, segment position (onset, medial, offset) do affect the variance of segment duration, and there're some directional invariance here, e.g., all segments with stress scale 3 are prolonged, two offset segments are prolonged also.

The duration of segment is constrained more decisively by the speaking speed. It's disclosed that the change rate of the segment duration of final (i.e. vowels, roughly) is more flexible than that of initial (i.e. consonant, roughly), it means that the durations of initial and final are not changed in the same rate while the speaking rate changed. And it's also shown that the location of a segment in the phrase influences the change rate of the duration also [6].

### 3.2. Linear Duration Model

The speech in the training database transcribed manually with C-ToBI, which had been used to build a Mandarin TTS system [7], was pronounced in reading style by a male speaker born at Beijing, with a normal speaking rate, and almost all sentences are indicative. So, the application domain of the duration model is fixed to the indicative sentence uttering with normal speed.

A linear segmental duration model was purposed for quantifying the input vector is consisted by the following, and here the segment means the phone.

1. segment ID (phone)
2. phone class index, initial classes and final classes
3. context in syllable, in phone class
4. number of syllable(s) in the word
5. number of word(s) in the intermediate phrase
6. number of intermediate phrase in the intonational phrase
7. break index, left boundary type for initial, right boundary type for final
8. stress index, syllable in the word
9. stress index, word in the intermediate phrase
10. stress index, intermediate phrase in the intonational phrase

The duration model is to map each index of each factor on a numerical value, and then combine these numerical values by multiplying them:

$$Duration(\vec{d}) = \prod_i S_i(d_i)$$

where  $d_i$  is the  $i$ -th element of vector  $\vec{d}$ .

The model is based on such an assumption that all factors have the property of directional invariance. And the parameter is estimated by holding all else constant and then got the statistical estimation.

## 4. CONCLUSION

C-ToBI in a hierarchical structure has been refined. The syllable is chosen as the basic unit for prosodic analysis. Intonational phrase, intermediate phrase, word, and syllable are the four layers of the prosodic structure tier. The break index transcribes the continuity of each pair of constituents in different layers, and the stress index transcribes the degree of accent of each constituent in different layers. The format of this kind of prosodic annotation can be easily used for the application. A linear duration model was built from a C-ToBI labeled database.

A pitch contour predictor based on the decision tree method had shown the positive effort of the prosodic model [7]. The next work can be expected is that a more complete prosodic model including duration and pitch contour predictor, which is built by more factors – phoneme and prosody from a C-ToBI labeled corpus.

## 5. REFERENCES

1. Santen J. "Prosodic Modeling in Text-to-Speech Synthesis", Proceedings of Eurospeech'97, Rhodes, Greece, 1997
2. Santen J., Shih C., Möbius B., Tzoukermann E., and Tanenblatt M. "Multi-Lingual Duration Modeling", Proceedings of Eurospeech'97, Rhodes, Greece, 1997
3. Silverman K., et al. "TOBI: A Standard for Labeling English Prosody", Proceedings of ICSLP'92, 1992
4. Tseng C., and Chou F. "A Prosodic Labeling System for Mandarin Speech Database", Proceedings of ICPhS'99, San Francisco, USA, 1999
5. Li Z. *Prosody Labeling System for Chinese: Report on CoSS (Corpus of Speech Synthesis)*, CDROM, Beijing, China, 1998, in Chinese
6. Zhang J., Zhu W., and Gårding E. "Temporal Structures of Spoken Chinese Sentences", Proceedings of ICPhS'99, San Francisco, USA, 1999
7. Niu X., Shen L., Zhu W., and Shi Q., "Modeling and Decision Tree Based Prediction of Pitch Contour in IBM Mandarin Speech Synthesis System", Proceedings of ISCSLP 2000, Beijing, China, 2000