Duration Prediction in Mandarin TTS System

*Qing GUO, Nobuyuki Katae**

Fujitsu Research and Develop Center China, Beijing, P.R.C
* Fujitsu Laboratories Ltd., Japan
guoqing@frdc.fujitsu.com

Abstract

This paper reports the methodology and results of decision tree based duration prediction for a Mandarin text-to-speech system developed by the Fujitsu Laboratories. Syllable initials and finals are the basic units in this duration study. Factors influencing finals duration such as phrase boundary and phone context are discussed in detail. Experiments indicate that it is the most important determinant of finals duration whether the prosodic factor of the right phrase boundary level is below the prosodic word level or not. Furthermore, the degree of phrase boundary vowel lengthening may vary depending on the types of finals. This paper also explains methods for objective evaluation of duration prediction model. Lastly, prosody evaluation results convincing that the prosody generated by our prosody generation module is much better than that of two other popular Mandarin TTS systems.

1. Introduction

Duration is one of the most important prosodic features contributing to the perceived naturalness of synthetic speech. Variation in segmental duration can hint at the identity of the speech sound and help segment a continuous flow of sounds into words and phrases thereby increasing naturalness and intelligibility. In natural speech, segmental durations are highly context dependent. Many contextual factors such as the phonetic identity of a current segment and its surrounding phone identities, phrase boundary level, phrasal position, and stressing or unstressing of segments can affect segmental duration substantially. The primary goal of duration prediction is to investigate the effect of these factors, so as to improve the naturalness of a text-to-speech system.

Many previous duration studies in Mandarin took the form of controlled experiments, where a limited number of contextual factors were examined in a small speech database and in a fixed sentence frame[1][2][3].

Chilin Shih et al [4] used a greedy algorithm to generate a small database which is rich in factors relevant to duration studies. They investigated the intrinsic scales of all categories of Mandarin phones and the major factors affecting their durations. In their paper, they reported the scales of vowel, fricative, burst-aspiration duration, and closure duration, and listed 14 factors influencing them respectively.

Based on a very large phonetic and prosodic enriched single speaker database, Min Chu et al [5] studied factors influencing durations of syllables in Mandarin. Six factors were investigated by comparing the durations in different categories. In this paper, it was reported that boundary index had the greatest influence on syllable durations, with tone identity ranking second.

Several segmental duration modeling methods have been widely used: additive model, multiplicative model, sums-of-product model[6] and the decision tree-based statistical (or CART) method. Since a very large phonetic and prosodic enriched database was available, Fujitsu Mandarin TTS system adopted decision tree based duration modeling.

This paper is organized as follows. Section 2 explains the corpus design and its transcription labeling. Section 3 discusses the results of the finals experiment. Section 4 gives the results of paired comparison of the prosody evaluation. Section 5 provides the conclusion.

2. Speech database

2.1. Basic synthesis unit

In the Fujitsu Mandarin TTS system, syllables are the basic synthesis units in the unit selection model. However, in order to avoid serious speech quality degradation, initials and finals are processed respectively with a PSOLA algorithm to have the speech prosody and concatenating speech waveforms modified.

There are about 205 syllable initials and finals in Mandarin and 1,600 tonal syllables in Mandarin. Using syllable initials and finals as basic units can improve the robustness of a statistic decision tree model. In our system, 182 finals and 23 initials have been defined in addition to the retroflex finals.

In our study, syllable initials and finals (with tone) are defined as the basic synthesis units.

2.2. Corpus design

Speech corpus design is critical in building high quality text to speech synthesis systems. Usually read speech is utilized, for it seems to be the easiest way to obtain a recorded speech corpus with highest control of the content. Comprehensive linguistic phenomena coverage is essential for a high quality synthesized speech TTS system. Greedy algorithm[7][8] was adopted for sentence selection to cover more phonetic context and prosodic context phenomena in a recording corpus of a given size.

The text source of our database was the Chinese People Daily 1998 Corpus, which is transcribed from a Chinese newspaper, with word segmentation and POS-tag annotated for natural language processing purpose.

In order to reduce the size of the feature space covered in the greedy algorithm, phone context and tone context have been classified respectively, much the same way that Min Chu et al [7] used. However, a little modification was made.

2.2.1. Feature space description

All tonal syllables that occur in the People Daily 1998 corpus are represented by their phone context and tone context information. In the 1998 People's Daily, there are 1,550 different tonal syllables. The size of the feature space is 1,550 (the number of tonal syllables) × 14 (the number of categories

of left phone context) × 22 (the number of categories of right phone context) × 3 (the number of categories of left tone context) × 3 (the number of categories of right tone context) = 4,296,600. However, not all these instances occur in real text.

There are 551,047 different vectors in the corpus totaling 22,596,405 occurrences in the corpus.

2.2.2. Speech corpus description

18,985 high frequency vectors were selected to cover about 50% of all occurrences. Another constraint was to include at least 5 different vectors for each tonal syllable. Using greedy algorithm, when 2,536 sentences were selected, all 18,985 high frequency vectors are covered.

In this set, 4,858 of 5k high frequency words were covered, and 8,482 of 10k high frequency words were covered.

Some of these 2,536 sentences are too long to pronounce as one single sentence. After segmenting those long sentences into readable sentences of appropriate length, 3,277 sentences were obtained.

83 sentences were selected to cover 119 high frequency r-syllables (with retroflex final in it) not covered by those 3,277 sentences. Eventually 3,360 sentences (with about 200k Chinese characters in them) and 1,550 isolated tonal syllables were recorded in our TTS corpus.

Boundaries of initials and finals were labeled with HTK toolkit and then checked manually.

2.3. Prosody structure and stress labeling

The prosody structure is composed of four tiers[9]: prosodic word (PW), minor phrase (MIP), major phrase (MAP) and intonation group (IG). Prosodic word is a tone group bearing one word stress. Minor phrase contains one or more prosodic words, bears one phrasal stress and the perceived break between MIPs is longer than that between PWs. MAP contains one or more PWs, bears one phrasal stress and the perceived break between MAPs is longer than that between MIPs. The criterion for prosody structure labeling is listening perception. Major phrases are often marked by commas with incomplete pitch resetting while intonation groups are marked by periods, quotation marks or semicolons with full pitch resetting.

Additionally, three levels of stress have been defined, namely the stressed, the normal and the neutralized.

The following is a sample transcription of a certain sentence in the speech corpus. "|", "‖", "|||" and "@" represent PW, MIP, MAP and IG in the transcription respectively. A syllable marked with "_H" means it is a stressed syllable, and a syllable marked with "_L" means it is a neutralized one.

8 月(ba1 yve4_H)/t | 2 0 日(er4 sh%2_H r%4_H)/t | 清晨(qing1_H chen2)/t ， ||| 一(yi1)/m 支(zh%1_H)/q 满载(man3 zai4_H)/v ‖ 锅 碗 瓢 盆(guo1_H wan3 piao2_H pen2)/l 、‖ 桌 椅(zhuo1_H yi3)/n 、‖ 调 料(tiao2_H liao4)/n 、‖ 发电机(fa1 dian4 ji1_H)/n ‖ 等(deng3)/u | 家当(jia1 dang4_H)/n 的(de5_L)/u ‖ 流动(liu2 dong4_H)/vn | 支前(zh%1_H qian2)/vn 车队(che1_H dui4)/n ‖ 从(cong2_H)/p 郑州(zheng4 zhou1_H)/ns | 出发(chu1 fa1_H)/v 了(le5_L)/y 。@

2.4. Sample questions of the decision tree

In the question set of the decision tree, there were two types of questions. One was the phone context question, and the other was the prosodic question. Here are some examples.

QS 'L_issaas' { "k-*","t-*","p-*" } whether the left phone is aspirated stops;

QS 'R_isnasc' { "*+n","*+m" } whether the right phone is nasal consonant;

QS "L_is1stTone" {a1-*, ai1-*,…, vn1-*} whether the left tone is the first tone;

QT "R_PhraseBoundary_2" {*+2} whether the segment is in the end of a prosodic word;

QT "R_PhraseBoundary_01" {*+0, *+1} whether the phrase boundary level of current segment is below prosodic word level.

The phrase boundaries can be divided into six levels. Level 0 means the segment is only in the middle of a word. Level 1 means the segment is in the word boundary. Level 2 means the segment is in the boundary of a prosodic word. Level 3 means the segment is in the boundary of a minor phrase. Level 4 means the segment is in the boundary of a major phrase. Level 5 means the segment is in the boundary of an intonation group.

There were no stress related questions in our duration study, because we did not expect the linguistic processing module to provide reliable stress information.

3. Results of the finals experiments

Figure 1 shows a sample duration decision tree for vowel "a1". Questions are asked on the nodes; the answer to a certain question is posed at a node which will lead to further nodes through selection of the appropriate branch. A terminal node with no branch is defined as a leaf node. In the leaf node level, the average duration of all sample segments will be the predicted duration. "Occ" means times of occurrence of segments in current leaf node. From the figure we can see that at the root of the tree, the question "R_PhraseBoundary_01" is asked. It means whether the phrase boundary level of current segment is below the prosodic word level or not. If the answer is yes then the question "L_isaas" will be asked till the leaf node is reached.

In this section, only experiment results of finals are discussed.

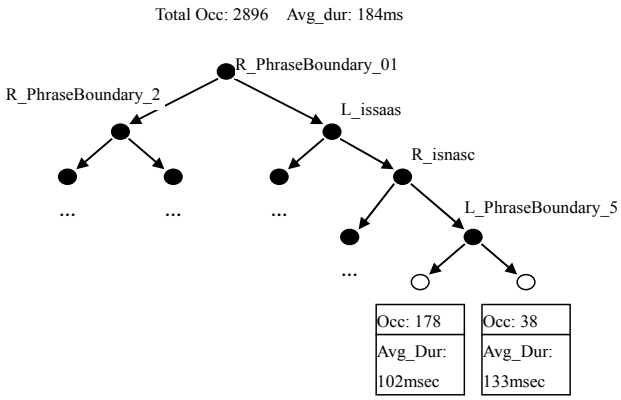

Figure 1. An example duration decision tree of vowel "a1"

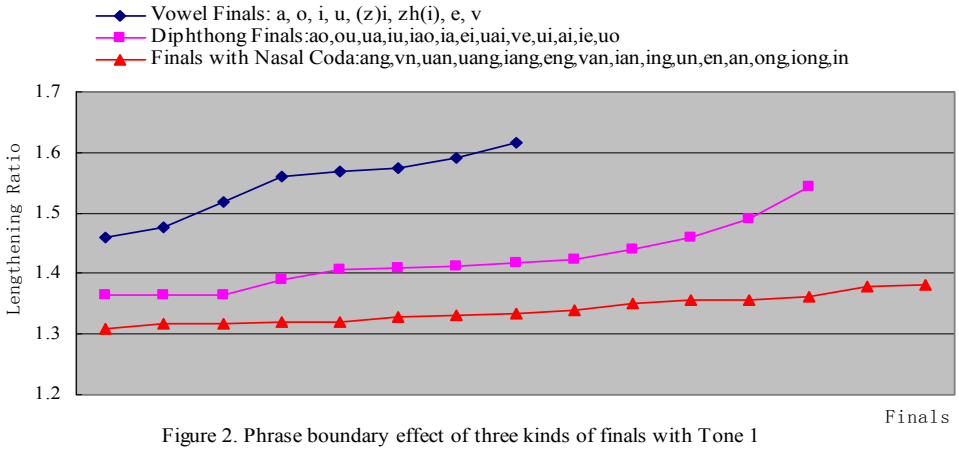

Figure 2. Phrase boundary effect of three kinds of finals with Tone 1

3.1. Right phrase boundary effect

For virtually all final decision trees, the first question is always to determine whether the right phrase boundary level of current segment is below the prosodic word level or not, i.e, R_PhraseBoundary_01. Some exceptions are finals with extremely sparse data. The decision trees of such finals have only a root node.

This means that the right phrase boundary level is the most important factor for final duration. Chilin Shih et al [4] and Min Chu et al [5] have also pointed out that the phrase boundary has significant influence on duration of vowels and syllables. Nevertheless, they used only one scale factor to describe to what degree the phrase boundary level influences duration of finals or syllables.

The average duration of a final with right phrase boundary level 0 (a final within a linguistic word) is almost the same as when it is in a right phrase boundary 1 (a final within a prosodic word but in a linguistic word boundary). To better measure the effect of phrase boundary, the ratio of the average duration of a final with phrase boundary 2, 3, 4 or 5 to the average duration of a final with phrase boundary 0 or 1, is used to describe the phrase boundary duration's lengthening effect. Meanwhile, the average duration of a final with phrase boundary 0 or 1 is regarded as the intrinsic scale of duration of the final.

Figure 2 shows that the right phrase boundary has a different duration lengthening effect on different finals with first tone. Generally speaking, for nucleus only (vowel) finals, phrase boundary has the greatest duration lengthening effect. With diphthong finals, the phrase boundary has a moderate duration lengthening effect. Finally with finals with nasal coda, the duration lengthening effect of phrase boundary is the weakest.

3.2. Tone identity

In many papers on Mandarin duration study, the identity of tone has been claimed to follow the property of:

Full tone > neutral tone; among full tones, 3 > 2 > 4 > 1

In our duration study, finals with neutral tone were found to have the shortest intrinsic scale of duration. But the other property, "among full tones, 3 > 2 > 4 > 1", was not supported by our experiment results. Most finals with the second tone were found to have relatively the longest duration among all the four full tones. This phenomenon will be further investigated in future studies.

3.3. Phone context and left phrase boundary level

Phone context questions were considered important factors that have great effect on the segment duration. Most previous duration research attempted to use influence coefficients to demonstrate how a certain phone context acts on the duration of current segment. However, after a thorough study with decision trees of different segments, we found it inappropriate to use the same scale to evaluate a phone context question's effect on duration of different segments.

For example, "R_l" and "R_isnasc" ({ *+n, *+m }) questions can be found in the decision trees for many finals, but not all finals. The "Yes" nodes of these two questions always bring shorter duration. As the 3 consonants (l, m and n) are voiced consonants, and the spectrum in the speech is continuous, we can assume that the finals before these 3 consonants should be uttered with a relatively shorter duration. However, the ratios of average duration of current finals with "R_l" and average duration of current finals without "R_l" vary in scales in different final decision trees.

Our large speech database features comprehensive phonetic context coverage. For example, the average sample number of each final is about 1,200, except for a few finals that have extremely low chance of appearance. A decision tree based duration model seems to be the more efficient method to grasp the various phone contexts and phrase boundary contexts.

Additionally, left phrase boundary level is also useful in the study of the duration of current final, though it is not a very important factor. For example, sometimes "L_PhraseBoundary_5" will be asked in the decision tree. "L_PhraseBoundary_5" means whether the final is in the beginning of a sentence or not. If yes, the duration of the final will be a little longer. This is consistent with the previous phonetic studies.

3.4. Evaluation

To evaluate the efficiency of a decision tree based duration model, the average relative predicted deviation is used to describe the accuracy of duration prediction. It is defined as follows:

$$avg\_dev = \frac{\sum_{sentence} \sum_{segment} \frac{|predict\_duration - real\_duration|}{real\_duration}}{total\_segments\_number} \quad (1)$$

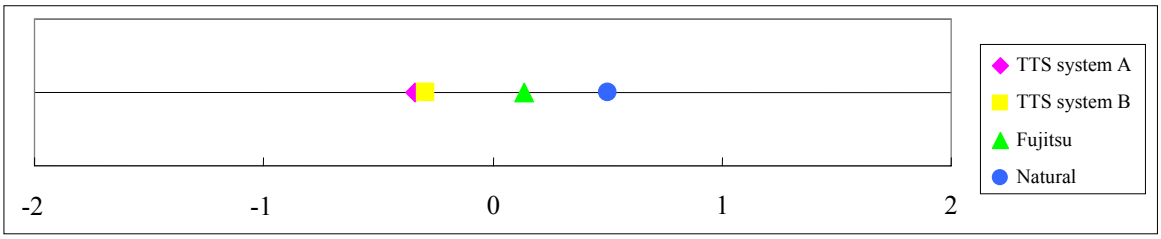

Figure 3.Scheffe Paired Comparison Test result of prosody evaluation

Two evaluation experiments were conducted. In the first experiment, the first 500 of the 3,360 sentences were not used to train the decision trees. Later they were used as a testing set to test the accumulated error of duration prediction. While in the second experiment, all 3,360 sentences were used to build decision trees, and these sentences were also used to test how effective the decision tree was. Table 1 gives the experiment results. The key difference between these two experiments is whether the testing set is included in the training set. From these experiments we can see that the duration decision trees based on large database are quite robust.

Table 1.Results of segment duration prediction accuracy

No.	avg_dev of finals	avg_dev of initials	avg_dev of all segments
1	14.8%	24.4%	19.6%
2	13.7%	20%	16.9%

4. Prosody evaluation

In order to evaluate the effectiveness of our prosody generation module (including both duration prediction and pitch prediction, the latter of which is not included in this paper), we conducted paired prosody comparing, in which prosody of our system, and that of two well-known Mandarin synthesized text-to-speech systems (referred to as system A and system B) were compared. We also compared our system with natural speech.

All of the waveform data are synthesized by a Fujitsu waveform generation module; therefore the only difference between the four kinds of synthesized speech is the prosody. In this way we were able to compare prosody extracted from natural speech (recorded by a female graduate student majoring in Chinese literature), prosody extracted from synthesized speech of the TTS system A, prosody extracted from synthesized speech of the TTS system B and predicted prosody using our prosody generation module.

10 sentences were used in the prosody evaluation. 10 people were asked to give their judgment on each pair of waveforms, A and B. There were 5 mutually exclusive choices for answers, namely: A is better; A is a little better; equal; B is a little better; and, B is better.

Figure 3 gives the results of the prosody evaluation. Not surprisingly, the natural prosody was rated the best. The prosody generated by our prosody generation module proved to be much better than that of two other popular Mandarin TTS systems.

5. Conclusions

In this paper, the factors influencing finals duration such as phrase boundary and phone context have been discussed in detail. Experiments indicate that it is the most important determinant of finals duration whether the prosodic factor of the right phrase boundary level is below the prosodic word level or not. Furthermore, the degree of phrase boundary vowel lengthening may vary depending on the different types of finals. Generally speaking, for nucleus only (vowel) finals, phrase boundary has the greatest duration lengthening effect. With diphthong finals, phrase boundary has a moderate duration lengthening effect. For finals with nasal coda, the duration lengthening effect of phrase boundary is the weakest. An objective measurement for evaluation of how well the duration prediction model works was also proposed.

Stress factor was not considered in our current duration study, for there is no reliable stress information available in present linguistic processing module. This factor will be studied in our future research.

Prosody evaluation results indicated that the prosody generated by our prosody generation module is much better than that of two popular Mandarin TTS systems.

6. References

[1] Feng, L., 1985. Duration of initials, finals and tones in Beijing Mandarin Speech. Acoustics Experiments in Beijing Mandarin, Beijing Univ. Press, pp. 131-195 (in Chinese).

[2] Cao, J.; Lu, S.; Yang, Y., 2000. Strategy and tactics on the enhancement of naturalness in Chinese TTS. Proc. International Symposium on Chinese Spoken Language Processing, Beijing.

[3] Zhu, W.; Matsui, K., 2000. A study of phoneme and syllable duration characteristics of Mandarin Chinese. Proc. International Symposium on Chinese Spoken Language Processing, Beijing.

[4] Shih C.; Ao B., 1997. Duration Study for the Bell Laboratories Mandarin Text-to-Speech System. Progress in Speech Synthesis, J. van Santen, R. Sproat, J. Olive, and J. Hirschberg, Eds. Springer, New York.

[5] Chu, M.; Feng, Y., 2001. Study on Factors Influencing Durations of Syllables in Mandarin. Proc. of Eurospeech, Scandinavia.

[6] van Santen, J.P.H., 1997. Prosodic modeling in text-to-speech synthesis. Proc. of Eurospeech, Rhodes, Greece, KN19-28.

[7] Min Chu; Hu Peng; Eric Chang, 2001. A concatenative Mandarin TTS system without prosody model and prosody modification. Proceedings of 4th ISCA workshop on speech synthesis, Scotland.

[8] Van Santen, J P. H.; Buchsbaum, A. L., 1997. Methods for optimal text selection. Proc. of Eurospeech, Rhodes, Greece, 553-556.

[9] Li A., 2002. Chinese Prosody and Prosodic Labeling of Spontaneous Speech. Prosody Speech, AIX-EN-PROVENCE France.