



# Improving Mandarin Segmental Duration Prediction with Automatically Extracted Syntax Features

Miaomiao WEN<sup>1</sup>, Miaomiao WANG<sup>1</sup>, Keikichi HIROSE<sup>2</sup>, Nobuaki MINEMATSU<sup>2</sup>

<sup>1</sup>Department of Electrical Engineering and Information Systems, the University of Tokyo, Japan

<sup>2</sup>Department of Information and Communication Engineering, the University of Tokyo, Japan

{wenm, wangm, hirose, mine}@gavo.t.u-tokyo.ac.jp

## Abstract

Previous researches have indicated the relevance between segmental duration and syntax information, but the usefulness of syntax features have not been thoroughly studied for predicting segmental duration. In this paper, we design two sets of syntax features to improve Mandarin phone and pause duration prediction respectively. Instead of using manually extracted syntax information as previous researches do, we acquire these syntax features from an automatic Chinese syntax parser. Results show that even though the automatically extracted syntax information has limited precision; it could still improve Mandarin segmental duration prediction.

**Index Terms:** speech synthesis, segmental duration, syntax feature

## 1. Introduction

Assigning appropriate segmental durations is important for the intelligibility and naturalness of text-to-speech systems. In some  $F_0$  generation systems [1], segmental durations of phones and pauses are first predicted and then used for the prediction of  $F_0$ -related parameters. So the precision of duration prediction is essential for the whole  $F_0$  generation process. Most Mandarin duration models rely upon phonetic and prosodic information in determining how long a phone or pause should last [2] [3]. While some other researches have investigated lexical or POS features, there are many new features to consider, especially syntax features. Though syntax features (except POS feature) have not been directly used in previous segmental duration prediction researches, [4] has studied the relation between syntax structure and pauses duration. Also, syntactic information has been widely used in predicting prosody structures of Mandarin [5].

It is necessary to predict segmental durations (including pauses) according to syntax information from the text. Firstly, this will help generation of prosody automatically at the back-end in a system. Secondly, from the view of human speech production, underlying and surface syntax representation of the utterance is the step before phonetic representation in human speech production process [6]. Syntax information might provide important cues for segmental duration prediction.

Manually parsing Chinese requires specialized knowledge and costs a lot of effort. To date, syntax information, such as syntax phrasal information and Subject-predicate relationship, used in previous researches on Mandarin prosodic segmentation prediction was manually extracted [4] [5]. POS feature is one kind of syntax information and is very commonly used in prosodic boundaries and segmental duration prediction tasks. This is, at least partially due to the reason that automatic POS

tagging has reached a high precision. Chinese syntax parser is far not perfect up to date. Therefore, rather than predicting durations entirely based on syntax structure, like in Japanese [1], we propose sets of syntax features to improve phones and pauses duration prediction. What is more, we acquire all the needed syntax information from an automatic syntax parser.

The data-driven approach of segmental duration prediction is realized mainly through establishing duration models, which build a mapping relationship between the information extracted from the text and the segmental duration. There are several methods to capture this relationship, including neural network [7], EM algorithm [8] and decision tree [9], etc. A polynomial regression model is proposed by [2]. As we do not focus on proposing a new model, we used two kinds of decision tree models, CART models and M5' tree model, to model a phone or pause in context. For all the modeling methods, how to select features is an essential problem. The syntax features we selected could be applied to other duration models.

The remainder of this paper is organized as follows. In Section 2, we describe the syntax parser and the syntax feature-sets we used. The baseline feature-sets and the decision tree models are described in Section 3. In Section 4, our experiments are described and the results are discussed. At last, this paper is concluded in Section 5.

## 2. Syntax features

### 2.1. Syntax parser

The parser we use is Stanford Parser version 1.6.2, whose performance is close to the best published figures for Chinese parsing [10]. For each input Chinese sentence the parser could output word segmentation, POS tagging, the phrase structure tree and grammatical relations (typed dependencies) of the sentence. Table 1 shows an output example.

### 2.2. Syntax features

In previous Mandarin duration prediction researches, prosodic features, which include contextual and structural characteristics of a speech segment, are used most often. Syntax features are much less frequently utilized in predicting durations. In addition to the prosodic and phonetic features, we extract two syntax feature-sets to improve phone and pause duration prediction respectively.

The syntax features are useful supplement for prosodic features. It is well known that Mandarin speech has four prosodic layers. For example, in the sentence below:

他#1一九三二年四月#2参加中国工农红军。

He joined the Chinese Workers' and Peasants' Red Army

| Phrase structure  |
|---|
| (IP<br>(NP<br>(DP (DT 这些–these))<br>(NP (NN 城市–city)))<br>(VP<br>(LCP<br>(QP (CD 三–three)<br>(CLP (M 年–year)))<br>(ADVP (AD 累计–collectively))<br>(VP (VV 完成–collectively)<br>(NP<br>(NP<br>(ADJP (JJ 固定–fixed))<br>(NP (NN 资产–asset)))<br>(NP (NN 投资–invest)))<br>(QP (CD 一百二十亿–12 billion)<br>(CLP (M 元–yuan))))))<br>(PU 。 )) |
| Typed dependencies  |
| det(城市–2, 这些–1), nsubj(完成–7, 城市–2)<br>nummod(年–4, 三–3), advmod(完成–6, 累计–5)<br>amod(资产–8, 固定–7), nn(投资–9, 资产–8)<br>dobj(完成–6, 投资–9), nummod(元–11, 一百二十亿–10)  |

Table 1: Stanford Parser output example. The translation of each word is added.

in April 1932.

#1 and #2 are both prosodic phrases boundaries. But #1 is also the subject-predicate boundary. #2 is time modifier boundary. A major pause is more likely to appear at #1 [4]. When predicting Mandarin prosodic information in the past, researchers usually only make use of the prosodic and phonetic information, neglecting much of the syntax information [2] [3]. Much contextual information is wasted. In this section we assume we could get syntax information with considerable accuracy from the parser and then describe the syntax features we use. Usually, the more various types of input information is used for estimation, the higher accuracy can be obtained. However, as size of the corpus is limited, too many kinds of input may cause a data sparseness problem. So we generally choose syntax features that are not only relevance to segmental duration but also have considerable occurrence rate.

### 2.2.1. Syntax features for phone duration prediction

Mandarin speakers naturally make some words more prominent in intonation than the other words within an intonational phrase. These words are said to be accented or to bear phrase accent. They are marked with changes in larger fundamental frequency range and longer duration. Perceptual experiments and acoustic studies showed that timing serves as the primary cue to the prominence. Also, the presence of prominence increases word duration [11]. Mandarin accent could be classified into two types, regular accent and emphatic accent. While emphatic accent is highly related to speaker’s intention and context, regular accent is the natural focus of a sentence read in neutral context. It is determined by the syntax structure of a sentence.

Here we will use syntax features to predict the regular accent and then add this accent information to phone duration predictor. There are many well-known rules for determining regu-

| Feature description                         | Parser output resource      |
|---|-----------------------------|
| Predicate in a subject-predicate structure  | nsubj<br>xsubj<br>nsubjpass |
| Modifier in a modification structure        | *mod<br>assm                |
| Object in a predicate-object structure      | dobj<br>pobj<br>lobj        |
| A location/person name or quantitative word | NR<br>CD                    |
| Regular Accent                              | –                           |
| Syntax depth of the current word            | Syntax tree                 |

Table 2: Syntax feature-set selected for phone duration prediction. The parser output resource lists from what parser output the feature is extracted. “\*” is the wildcard matching all possible prefix of the typed dependency name.

lar accent, such as predicate being the regular accent in Subject-predicate structure, object being the regular accent in Predicate-object phrase and modifier being the regular accent in Modifier-object phrase [11]. Also, the more information conveyed by a word, the more likely it will be a regular accent [11] [12]. These four rules have been experimentally verified by [12]. Therefore, the first four features of the feature-set (Table 2) are chosen according to these four rules respectively. Then another binary feature (Regular Accent) is used to indicate if a word is regular accented or not, according to the above four rules. The depth of a word in the syntax tree (syntax depth) is also added to the feature-set.

### 2.2.2. Syntax features for pause duration prediction

A well known assumption is that the length of pause is related to the grammatical constraints such as tightness of the two domains. Table 3 shows the syntax feature-set we used in pause duration prediction. These features are chosen mainly according to [4], which analysis the pause distribution among various syntax boundaries. Then, an empirical Pause Level feature is assigned to each prosodic-word boundary (Table 4), larger Pause Level value indicates longer pause. The Regular Accent of the previous and next word is also added.

## 3. Segmental Duration Prediction

Mandarin phones could generally be divided into two groups: initials and finals. The initials (21 phones in total) are consonants and the finals (35 phones in total) contain at least one vowel. As factors that influence initial duration are different from those for final duration, in our experiments, the initial and final durations are modeled separately.

### 3.1. Baseline feature-sets

In order to valid the usefulness of syntax features, the baselines for our experiments are obtained using only prosodic, phonetic and POS features, as listed below.

For initial duration prediction, the baseline feature-set includes: initial name, initial category, final of the current syllable, final category of the current syllable, tone of the current syllable, tone of the syllable before and after the current syllable.

| Feature description   | Parser output resource    |
|---|---------------------------|
| Is subject-predicate boundary or not  | nsubj, xsubj<br>nsubjpass |
| Is modifier boundary or not   | *mod                      |
| Is word boundary in word compound or not  | NP                        |
| After auxiliary word “的”  | assm                      |
| Is preposition-object or verb-object boundary or between linking verb and predicative | *obj<br>cop<br>attr       |
| Syntax depth of previous word   | Syntax tree               |
| Regular Accent of previous and next word  | –                         |

Table 3: Syntax feature-set selected for pause duration prediction. The parser output resource lists from what parser output the feature is extracted. “\*” is the wildcard matching all possible prefix of the typed dependency name.

| Boundary Type  | Pause Level |
|--|-------------|
| Subject-predicate  | 4           |
| Preposition-object or verb-object boundary or between linking verb and predicative | 2.5         |
| Modifier-object  | 2           |
| Words boundary in word compound  | 1           |
| After auxiliary word “的”   | 1           |
| Other  | 0           |

Table 4: Definition of Pause Level feature.

ble, prosodic boundary type of the boundary before and after the current syllable, syllable number of current word foot, syllable number of current prosodic word, syllable number of current prosodic phrase, syllable number of current breath group, POS of current word, syllable position in current word foot/prosodic word/prosodic phrase/breath group/sentence.

For final duration prediction, the baseline feature-set includes: final name, final category, initial of the current syllable, initial category of the current syllable, tone of the current syllable, tone of the syllable before and after the current syllable, prosodic boundary type of the boundary before and after the current syllable, syllable number of current word foot, syllable number of current prosodic word, syllable number of current prosodic phrase, syllable number of current breath group, POS of current word, syllable position in current word foot/prosodic word/prosodic phrase/breath group/sentence.

For pause duration prediction, the baseline feature-set includes: the previous syllable’s final category, the next syllable’s initial category, prosodic boundary type, syllable number of word before and after the pause, position in prosodic word, position in prosodic phrase, position in breath group, position in sentence, tone of the syllable before and after the pause, POS of previous and next word.

### 3.2. Decision tree models

We use two kinds of decision tree models in our experiments. One is the well known CART models [13]. We used Wagon, which is part of the Edinburgh Speech Tools Library. The other

|          | Train | Test | Total |
|----------|-------|------|-------|
| Initials | 6769  | 390  | 7159  |
| Finals   | 8026  | 450  | 8476  |
| Pauses   | 3066  | 168  | 3234  |

Table 5: The number of phones and pauses

one is M5’ tree model [14]. The M5’ can be used as a regression tree (M5p-R) or as a model tree (M5p). If a leaf, in M5’ algorithm’s building process, is associated with an average output value of the instances sorted down to it, then the model is called regression tree [14]. If the tree concludes in its leaves to more complex regression functions of the input variables, then the model is called model tree [15]. In dealing with the continuous features or predicting continuous value, CART program builds regression trees that differ from decision trees only in having values rather than classes at the leaves. Because we want to precisely predict the exact values of durations of pauses, initials and finals durations. M5’ model, which aims at dealing with continuous classes, is expected to obtain better prediction results.

## 4. Experiments and Results

### 4.1. Speech corpus and preprocess

Our speech corpus contains 300 sentences read by a native female speaker, arranged at University of Science and Technology of China. It includes pronunciation symbols and prosodic boundaries labels. The sentences are first word segmented by Stanford Chinese Word Segmenter [16], then all of them could be successfully parsed by the Factored Chinese version of Stanford Parser [10].

The 300 sentences are divided into train (90%) and test (10%) sets. As prosodic-word is defined as a group of syllables that should be uttered closely and continuously, any inner prosodic-word break will make the speech unintelligible or unnatural [5]. Thus we assume a short pause between two adjacent prosodic-words. Breath group boundaries always contain a long silence which is out of the research in the paper. Table 5 gives the statistics of the corpus.

### 4.2. Results and discussion

Table 6 summarizes results for prediction of initials and finals durations. The baseline and syntax feature-set is referred to as *B* and *S*, respectively. i.e., M5p-R-*B+S* stands for M5p-R tree model using both baseline and syntax feature-sets while M5p-R-*B* stands for M5p-R tree model using only baseline feature-set. All models using syntax feature-set perform no worse than their baselines in terms of both correlation and RMSE. Also, M5p-R and M5p outperform Wagon in terms of both correlation and RMSE. In particular, M5p-R-*B+S* is the model with best performance in both initials and finals duration prediction tasks.

Table 7 presents the results for pauses. When syntax features are added, the prediction results of both Wagon and two M5p models outperform their baselines. Also, M5p-R and M5p outperform Wagon in terms of both correlation and RMSE. In particular, M5p-*B+S* is the model with best performance in our pauses prediction task.

The results above indicate that syntax features could steadily improve the prediction of pauses, initials and finals du-

| Model             | Initials    |           | Finals      |           |
|-------------------|-------------|-----------|-------------|-----------|
|                   | Corr.       | RMSE (ms) | Corr.       | RMSE (ms) |
| Wagon- <i>B</i>   | 0.94        | 13        | 0.73        | 29        |
| Wagon- <i>B+S</i> | 0.95        | 13        | 0.76        | 27        |
| M5p- <i>R-B</i>   | 0.95        | 13        | 0.83        | 18        |
| M5p- <i>R-B+S</i> | <b>0.97</b> | <b>12</b> | <b>0.84</b> | <b>16</b> |
| M5p- <i>B</i>     | 0.95        | 13        | 0.79        | 25        |
| M5p- <i>B+S</i>   | 0.97        | 12        | 0.81        | 24        |

Table 6: Results for phone duration prediction. The maximum correlation/minimum RMSE values are shown in bold.

| Model             | Correlation | RMSE(ms)  |
|-------------------|-------------|-----------|
| Wagon- <i>B</i>   | 0.65        | 22        |
| Wagon- <i>B+S</i> | 0.67        | 20        |
| M5p- <i>R-B</i>   | 0.66        | 20        |
| M5p- <i>R-B+S</i> | 0.67        | 18        |
| M5p- <i>B</i>     | 0.67        | 20        |
| M5p- <i>B+S</i>   | <b>0.71</b> | <b>17</b> |

Table 7: Results of pauses duration prediction. The maximum correlation/minimum RMSE values are shown in bold.

rations. The improvement is not significant for initials. There are several reasons. First, this may due to that normal accent influence more on finals. Also, our training data contains only 270 sentences so the data sparse problem might be quite significant. Though direct comparison may not be very reasonable, we achieved better initial and final duration prediction results compared to [2], whose training set includes 10000 sentences with the same prosodic tagging scheme as ours.

To further testify the usefulness of syntax information in the segmental duration prediction task, and also to look more specifically at which of the features in our whole feature-sets are most useful, we performed feature selection using a selection algorithm that computed individual predictive power of each feature and the redundancies between features [17]. Table 8 shows the best five features for initial, final and pause duration prediction. The feature selection results reconfirm the usefulness of syntax features, especially for final and pause duration prediction.

## 5. Conclusion

In this paper, we have introduced the syntax features extracted from Stanford parser output, and demonstrated their usefulness in Mandarin phones and pauses duration prediction. The result is encouraging in that it shows that even though the Chinese syntax parser is far not perfect, the automatically generated syntax features could still improve segmental duration prediction by providing high level linguistic information.

In our future work, we will try to combine the output of different Chinese parsers to improve the quality of syntax features. Also, we will further testify the usefulness of syntax information in Mandarin prosody modeling.

## 6. Acknowledgements

The authors of this paper would like to thank Prof. Renhua Wang in the University of Science and Technology of China for

| Rank | Initials                    | Finals                  | Pauses  |
|------|-----------------------------|-------------------------|---|
| 1    | Initial name                | Final name              | The next syllable's initial category            |
| 2    | Position in prosodic-phrase | Tone                    | Position in sentence                            |
| 3    | Position in prosodic-word   | Syllable number of word | <b>Is word boundary in word compound or not</b> |
| 4    | Initial category            | <b>Regular Accent</b>   | <b>Pause Level</b>                              |

Table 8: Best four features for predicting initial, final and pause duration. The syntax features are shown in bold.

offering us the speech corpus.

## 7. References

- [1] K. Hirose *et al.*, "Corpus-based generation of prosodic features from text based on generation process model", Proc. Interspeech, pp.1274-1277, 2007.
- [2] S.Lu *et al.*, "Polynomial regression model for duration prediction in Mandarin", ICSLP2004, pp.777-780, 2004.
- [3] Jian Yu *et al.*, "The Pause Duration Prediction for mandarin Text-to-Speech System", IEEE NLP-KE 2005, pp.204-208, 2005.
- [4] J. Cao *et al.*, "Syntax and Lexical Constraint in Prosodic Segmentation and Grouping", Speech Prosody 2002.
- [5] Chu M. *et al.*, "Locating Boundaries for Prosodic Constituents in Unrestricted Mandarin Texts". Computational Linguistics and Chinese Language Processing, February 2001, Vol.6.No.1:pp.61-82,2001
- [6] Cooper William *et al.*, "Syntax and speech", Cambridge, Mass.: Harvard University Press, pp.2-3, 1980
- [7] S.H. Chen *et al.*, "An ANN-based prosodic information synthesizer for Mandarin text-to-speech", IEEE trans. Speech Audio Processing, Vol.6, No.3, pp.226-239, 1998.
- [8] S.H. Chen *et al.*, "A new duration modeling approach for Mandarin speech", IEEE trans. Speech Audio Processing, Vol.11, No.4, pp.308-320, 1998.
- [9] Hyunsong Chung, "Duration models and perceptual evaluation of spoken Korean", Speech Prosody 2002.
- [10] <http://nlp.stanford.edu/software/lex-parser.shtml>
- [11] Qian, Y. *et al.*, "Assigning Phrase Accent to Chinese Text-to-speech System", International Conference on Acoustics, Speech, and Signal Processing, Vol.1, pp.485-488, 2002.
- [12] Lu Shinan *et al.*, "Prosodic control in Chinese TTS system", IC-SLP2000, Vol.1,pp.21-24, 2000.
- [13] Black, A. *et al.*, "Edinburgh Speech Tools Library: system documentation. Technical Report 1.2.0 edition", The Centre for Speech Technology Research, University of Edinburgh, UK.
- [14] Quinlan, R.J., "Learning with continuous classes", Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, Nov. 16-18, World Scientific Pub Co Inc., pp.343-348, 1992.
- [15] R Wang *et al.*, "Induction of model trees for predicting continuous classes", Proceeding of the Poster Papers of the European Conference on Machine Learning, Springer, Prague, Czech Republic, pp.128-137, 1997.
- [16] <http://nlp.stanford.edu/software/segmenter.shtml>
- [17] Witten I.H. *et al.*, "Weka: Practical machine learning tools and techniques with Java implementations", ICONIP/ANZIIS/ANNES' 99 International Workshop, Dunedin, 1999.