Modeling Duration and Intonation in Mandarin Chinese Synthesis with a Neural Network

Hongwei DING, Oliver JOKISCH, Hans KRUSCHKE
Laboratory of Acoustics and Speech Communication
Dresden University of Technology, Dresden, Germany
{hongwei.ding, oliver.jokisch, hans.kruschke}@ias.et.tu-dresden.de

ABSTRACT

The prosody control plays an important role in the naturalness of synthesized speech. In previous work, great efforts have been made to generate rule-based or parameter-based prosodic models [6]. In order to capture the complex interaction of different relevant prosodic factors, neural networks were recently employed. This paper presents a new method of learning and modeling duration and intonation in Mandarin Chinese synthesis with a neural network, which was proved to be an appropriate approach in our Mandarin synthesis system. The material for the study of prosodic components was extracted from a phonetically and prosodically labeled sentence database uttered by the same speaker as for the synthesis inventory. This paper reports the study of duration and intonation, the analysis of the database, the concept of neural network model and the evaluation of training results.

1. INTRODUCTION

A good pitch prediction in a text-to-speech system is essential for intelligibility and especially important for the natural sounding of tonal languages such as Chinese. The significance of the prosodic modeling in Mandarin Chinese has also been revealed by many recent publications found in this field.

- In the duration study, instead of examining performance in controlled conditions, where limited number of factors were allowed, large speech corpus with balanced phonetic and prosodic coverage was preferred: Zhu conducted duration study with C-ToBI labeled corpus [1]; Chu investigated syllable duration with regard to many influencing factors [2].
- In the intonation generation, prosodic models tried to build up the pitch movement in prosodic words as well as the base intonation contour: Hu was successful in prosody generation using the template of quantified prosodic unit and base intonation contour [3]. Neural networks (NN) were also introduced in this field. Chen proposed to generate prosodic information from base syllable [4]. Chen described a multi-strategy data mining framework to generate the fundamental frequency and energy contours [5].

By incorporating the advantages of the duration study with natural speech database and superposition of tonal coarticulation with sentence intonation from the above models, with consideration of our special designed inventory, we propose a new approach in this paper to generate prosody with neural network.

Since our TTS database was designed for TD-PSOLA synthesizer, we decided to integrate as much as possible signal variations into the basic speech units. The variety of inherent tone contours and the coarticulatory effect across syllable boundaries were built in the inventory. Thus the acoustic inventory includes all 4 tonal variations as well as contextual coarticulatory variations of each syllable, which amounts to 3,049 units. In order to design a suitable prosodic model for our synthesizer, a new neural network is suggested, which uses the linguistic and prosodic information to predict duration and intonation efficiently.

This paper will first discuss the duration and intonation study in section 2. In section 3 database collection and analysis will be described. The neural network will be introduced in section 4 to train these data and predict the prosody. The results are presented in section 5, which proved that the NN is efficient to realize our goal of prosody generation.

2. INTONATION AND DURATION STUDY

2.1 Intonation model

In one of our previous studies [6], the intonation model was the superposition of suprasegmental sentence intonation, and segmental intonation component. This theory is also reflected in this research, but the neural network makes it possible to treat these in combined effects. One synchronization point in the middle of the syllable was chosen for each tonal change due to sentence intonation effect. In order to model the tone,
coarticulation effects, two points were calculated at the onset and offset of the finals of the tone to catch the peaks and lows in the syllable contours, and for the sake of tonal coarticulation. One point at the beginning of the initials was also introduced, thus we have four points in the syllable, which can capture the maximum and minimum of the tone contours, and follow the tonal course.

An experiment was conducted to investigate whether these four points can generate the intonation for the synthesized sentences properly. The measurements of the points were obtained from graphical plots of pitch tracks with an interactive interpolated sentence intonation contour. Figure 1 at the top is one example to show the efficiency of this assumption.

The cross points show the pitch movement from the synthesized phrase “qing4 zhu4 zhong1 guo2 de5 niu2 nian2”, which resemble the original pitch contours described in lines. Both “qing4” and “zhu4” are Tone 4, they have similar pitch value from inventory, and when the pitch manipulation is applied, the first Tone 4 does not drop to the end, and second Tone 4 begins not so high, so that the sentence intonation and tonal coarticulation effects are very well reflected. And these effects can also be observed at the smooth connection between the latter part of the phrase.

So we can draw the conclusion, that the four points in a syllable can model not only the sentence declination but also the tonal coarticulation. So that if the neural network can provide the accurate measurements of these four points, a good intonation can be expected for the synthesizer.

2.2 Duration study

In previous investigation [6], a multiplicative model was adopted, which is based on a multiplicative combination of the inherent phoneme duration values and contextual influence factors. The duration study in this paper is however based on syllable. The important factors include:

- The actual syllable and its lexical tone
- Syllable type and tone of the preceding syllable
- Syllable type and tone of the following syllable
- Number of preceding and following syllables in the word
- Number of preceding and following syllables in the phrase
- Accentuation

The network will learn the interactive influence of these factors and predict their performance in the synthesis.

3. MATERIAL COLLECTION AND DATABASE ANALYSIS

3.1 Material collection

The prosody in natural speech is effected by speaker-dependent factors, and non-speaker dependent factors. In order to facilitate the prosodic modeling in time domain speech, the database is prepared from the inventory speaker.

The natural speech material of the text database was selected from newspaper articles. The style of the material is mostly newspaper 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 various parameters which have influences on the duration of the syllable. 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 text database includes 63 sentences ranging from 15 to 86 syllables per sentence, which amounts total 2385 syllables.
3.2 Database analysis

In order to provide enough information for the duration and intonation study, the labeling includes phonetic and prosodic labels at several layers:

- The phonetic label includes initials, finals and lexical tones. A labeling table of 73 tokens consists the individual sounds and 5 tones. The closure and release part was labeled separately.
- The prosodic label includes the prominence levels of the syllable (phrase- and word-stress), position of the phoneme (boundary of syllable, word, phrase or sentence).

Since the label information supplies the neural network with the input to learn, reliability is necessary. Hence the labeling was conducted manually.

Besides the above mentioned phonetic and prosodic information, some other information such as syllable type has also been collected. The labeling was progressed onto the level of phonemes, the learning and training of neural network was, however, based on syllable. A syllable-oriented table of representation was obtained. One excerpt was taken from the table in Figure 2, the syllable (syl) is accompanied with:

- Syllable type (syltyp)
- Its lexical tone (tone), previous tone (prvtone), and next tone (nxttone)
- Prosody word information: number of syllables in word (nsyl/w); number of words (nwrds), number of syllables (nsyl), number of phrases (nphr) in sentence.
- F0 is measured at the beginning of the initial consonants (f0-c-left), at the beginning (f0-v-begin), in the middle (f0-v-middle) and at the end of the final vowels (f0-v-end).

- Other information includes: the sentence mode (snttype), the sum of mean phoneme duration values from the statistics (sumdur). Sylldur and paudur are the original syllable duration and pause duration, which are the learn targets

The database was especially designed for the training task and is much smaller than a corpus database. Since the complete labeling and statistics were carefully manually processed, it contains very reliable information for the network training. Of course, LNRE (large number of rare events) problems are not covered by this database.

### 4. NEURAL NETWORK

Neural networks represent one adequate method to learn the intended prosodic contours and parameters. 17 linguistic and phonetic input features have been selected from the prosodic database already described in the following order: isyl, isyltyp, tone, syl/wrd, nsyl/w, wrd/phr, wrd/syl, phr/syl, syl/phr, wstress, pstress, prvtone, nxttone, nwrds, nsyl, nphr and sumdur. The learn targets are also defined by the database: sylldur, paudur and the 4 f0 model points, respectively.

![Fig.3a: Training of the prosodic model.](image)

Figure 3a and 3b show the use of the neural network in training and test phase.
5. RESULTS

In order to evaluate the results of the novel prosodic control, perceptual experiments were carried out. The prosodic modifications in resynthesis and synthesis were TD-PSOLA based.

In the resynthesis, the undesired effects caused by the acoustic units modified by PSOLA can be neglected. The evaluation of resynthesis sentences with a MOS of about 4.8 (natural sentence should be 5.0) reveals that the prosodic prediction from neural network is appropriate.

Quality improvement has also been achieved in synthesized sentences. Synthesized sentences without any prosodic modification normally achieve a MOS impression of 1.90. If the neural network prosodic modification applied, a general quality MOS of 3.39 was obtained.

6. DISCUSSION

Using a small but especially designed syllable database and enhanced linguistic feature set, a neural network approach enables an efficient prosody control in Mandarin Chinese concatenative speech synthesis. Some improvements can still be made in our approach. For example, because Mandarin native speakers are very sensitive to every pitch movement in the tone, more points can be selected in the tone contour to follow the pitch track more accurately. Furthermore all the necessary linguistic features should be generated automatically from text analysis.

Another concern is that one can not expect the output of a NN to behave well in a situation out of its training set, the corpus will be enlarged in the future.

7. LITERATURE