LEARNING SYLLABLE DURATION AND INTONATION OF MANDARIN CHINESE

Oliver Jokisch, Hongwei Ding, Hans Kruschke and Guntram Strecha

Dresden University of Technology
Laboratory of Acoustics and Speech Communication, 01062 Dresden, Germany
{oliver.jokisch,hongwei.ding,hans.kruschke,guntram.strecha}@ias.et.tu-dresden.de

ABSTRACT

The perceived quality of synthetic speech strongly depends on its prosodic naturalness. The current paper presents a neural network based prosody model of Mandarin Chinese. Using a small but especially designed syllable database and an enhanced linguistic feature set, the novel approach enables the training of syllable duration, syllable-based F0 model points and is suitable for the multilingual prosody control in concatenative speech synthesis. The paper describes database design, neural network model, training results for Chinese and the perceptual evaluation. The results indicate the importance of the appropriate database design and the enhanced linguistic feature set. Perceptual tests of resynthesized stimuli using predicted duration values receive MOS comparable to natural speech of about 4.8.

1. INTRODUCTION

The synthesis of near-to-natural prosodic contours is still an important issue in text-to-speech (TTS). Recent TTS systems concatenating larger speech units or variants from a corpus achieve a considerably high quality, because they preserve the natural prosodic structure at least throughout the units. Though, these systems are often domain-specific and require a larger amount of collected data. An optimal unit-selection still demands improved prosodic models.

Data driven algorithms for prosody control enable the simple adjustment of prosodic parameters by training and lead to more variable and natural contours. The multilingual TTS system DRESS includes a Mandarin Chinese, male voice basing on the concatenation of 3.049 syllables with lexical tones. The prosodic features of Chinese are predicted by a novel neural network according to the design in [1]. Recent examinations of German as in [2] underline the importance of enhanced linguistic and phonetic input information. Including 24 different input features resynthesis stimuli of network-predicted syllable duration achieved a remarkable Mean Opinion Score (MOS) of 4.08 versus a MOS of 4.51 testing natural speech stimuli. The current research focuses on the selection of appropriate linguistic and phonetic features and the database design for the neural network training. The study includes perceptual experiments.

Great efforts have been made to improve the prosody of Mandarin Chinese in the past decades. Duration factors were studied [3]. Rule-based and parameter-based models have been suggested in a few TTS systems. Due to the advantages of data driven models, it has also been suggested to automatically predict the prosodic phrase boundaries [4]. A RNN-based prosodic information synthesizer has also been managed for Mandarin Text-to-Speech [5]. It generates prosodic information, including the pitch contour, etc. on the basis of 411 base syllables. 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. Therefore a neural network was designed to generate prosodic information with the assumptions that:

1. The variety of inherent tone contours should be preserved as much as possible.
2. The coarticulation effect at the cross syllable boundaries will be considered.

The acoustic inventory includes all 4 tonal variations together with some additional neutral tone variants, which amounts to 3.049 units. The neural network generates the information for both syllable duration values and intonation modification information. With regard to the perceptual evaluation this integrated neural network prosodic model performs very well for the Mandarin Chinese synthesis.

2. CHINESE DATABASE

2.1. Concept and data collection

In contrast to rule-based, parametric prosodic models [6, 7], data-driven systems have some advantages, because it is difficult to identify all prosodic factors and to model their interaction in natural speech. A neural network approach can catch the behavior by training sufficient amounts of appropriately labeled data. A careful preparation of the database will help to learn the influencing prosodic factors.

```
isent isyl   syltyp    tone  nsyl/w   wrd/sen  syl/phr pstress nxttone nsyl  snttype     syldur     f0modelpoint
  sent   syl    isyltyp   syl/wrd wrd/phr phr/sen  wstress prvtone nwrd   nphr   sumdur       paudur

0 1006 0 jiu4  CVV   10  4   1   2   1   1   1   1   1   0   0   4  34  62  9  1  224.6    134    0    205.1
0 1006 1 suan4 CVVN  12  4   2   2   1   1   1   2   0   1   4   0  34  62  9  1  281.9    266    0    169.8
0 1006 2 pause pause  1  0   0   0   0   0   1   0   0   0   4   4  34  62  9  1  485.6      0  505    195.8
0 1006 2 z14   CV     3  4   1   2   2   2   1   3   1   0   0   3  34  62  9  1  188.4    164    0    240.6
0 1006 3 ji3   CV     3  3   2   2   2   2   1   4   0   0   4   0  34  62  9  1  179.9    112    0    139.2
```

Fig. 1. Excerpt from the prosodic syllable database of Chinese

On the other hand, the prosody in natural speech is effected by speaker-dependent and speaker-independent factors. In order to facilitate the prosodic modeling in speech synthesis, the prosodic database is also prepared from the inventory speaker. Two kinds of material were chosen for the analysis of prosody:

- Phonetically-labeled speech inventory: The syllables were segmented from carrier sentences [8], in which all the 3049 target syllables were followed by neutral tones. The syllables keep the inherent prosody and their tone contours are least influenced from the following tones. These characteristics proved to be important for time domain synthesis.
- Prosodically-labeled natural read texts: The text database (63 sentences of 15 to 86 syllables, total 2385 syllables) was selected from newspaper articles (mostly news reports). The speaker was asked to read the texts at a normal speed in a fluent way. The sentences were selected from recordings of approximately one hour.

2.2. Prosodic analysis

The initial step of the database analysis is to decide on the important prosodic factors in synthesized speech. So called *core factors* (phone identity, stress-related, locational and tone-related) have been already defined in [9].

To cover these factors, the database labeling was carried out in phonetic and prosodic categories at the same time. Phonetic labeling includes the label of initials, finals and the lexical tones. A labeling table of 73 units consists of individual sound segments together with five lexical tones. The closure and release part of the plosives were labeled separately. Different levels of pauses, between sentences, prosodic words and phrases were also indicated. The prosodic labels include the prominence levels for syllables (phrase- and word-stress), the phoneme position (boundary indication: syllable, word, phrase, sentence) and the phrase mode (declaration, question, etc.). The labeling was conducted manually, so that the accuracy can be guaranteed in contrast to automatic labeling. The labeling was processed in three layers: the prominence information, the location information and the information of acoustic segment and its lexical tones. All the relevant information concerning the previous and the next segment could also be obtained in the analysis.

Besides the *core factors*, some other information like the syllable type has been collected. The statistics was conducted on the level of syllable and phoneme. Finally, a syllable-oriented table of representation was obtained. For the table excerpt shown in figure 1, syllable (syl) *jiu4* belongs to syllable type (syltyp) *CVV (consonant-vowel-vowel)*, occurring in *tone* 4. It is the first syllable in the word (*syl/wrd=1*), which is a disyllabic word (*nsyl/w=2*) found in the first word of the phrase (*wrd/phr=1*) and in the first word of the sentence (*wrd/sen=1*). It appears in the first phrase of the sentence (*phr/sen=1*) and in the first syllable of the phrase (*syl/phr=1*). The syllable has word-stress (*wstress=1*), but no phrase stress (*pstress=0*). It is preceded by silence (*prvtone=0*), followed by tone 4 (*nxttone=4*). There are some more information about the sentence, in which the syllable appears: *nwrd=34*, *nsyl=62*, *nphr=9*, *snttype=1* mean that there are 34 words, 62 syllables, 9 phrases in the sentence and a declarative sentence mode, respectively. *sumdur* is the sum of mean phoneme duration values from the statistics. *syldur* and *paudur* are the original syllable duration and pause duration (learn targets). The fundamental frequency (f0) is measured at the left margin, in the center and at the right margin of the syllable final. (*f0left*, *f0cent*, *f0right*). For network training only a single f0 value (*f0 model point*) represents each syllable (*f0left* for tone 3; else *f0cent*). 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.

Because each sound unit in the inventory carries sentence stress, the average phoneme duration in the inventory is 147 ms versus an average duration of 105 ms in the read texts. In the course of prosodic modification, the duration of plosive consonants will not be modified. Without consideration of plosives the average phoneme duration is 159 ms (inventory) and 113 ms (prosodic database), respectively.

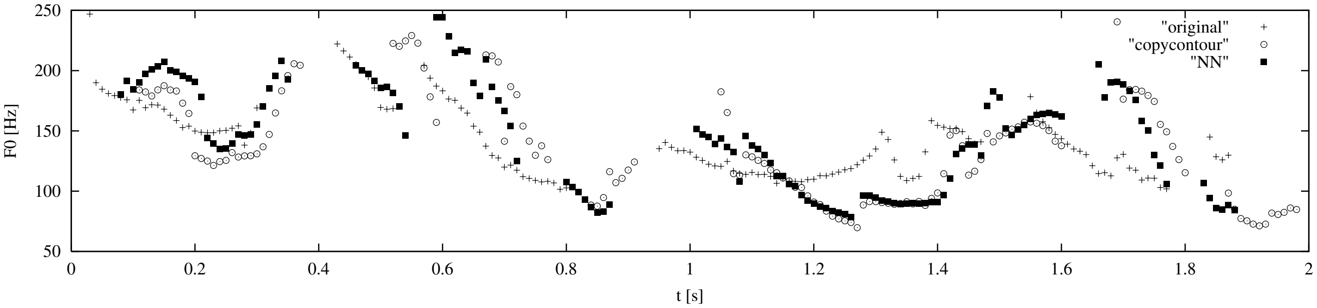

Fig. 2. Measured *f0*-contours: original vs. synthesis with original f0 (*copy contour*) and network-predicted f0 values. Excerpt from the utterance ”gui1 mo2 zui4 da4 de0 shi4 mei3 nian2 yi1 du4 de3 ...” (The greatest yearly event is ...).

3. NEURAL NETWORK TRAINING

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 described in section 2 in the following order: *isyl, isyltyp, tone, syl/wrd, nsyl/w, wrd/phr, wrd/sen, phr/sen, syl/phr, wstress, pstress, prvtone, nxttone, nwrd, nsyl, nphr, sumdur*. *sumdur* characterizes the sum of the mean phoneme duration values in the particular syllables obtained from the database and can be considered as a type of *base duration*. The learn targets are also defined by the database: *syldur, paudur* and *f0 model point*, respectively.

According to [2] the training and prediction tasks are solved by a fully-connected feed-forward neural network (FFNN) of three layers (17x10x3 neurons). Depending on the parameter ranges the input and output parameters are linearly scaled. Both log and tan-hyperbolic transfer functions are used. The FFNN is trained using a teaching input (prosodic database) and standard error backpropagation, minimizing the root mean square error (RMSE) between teaching input and network output. The database (63 sentences, 2385 syllables) was subdivided into a training set (50 sentences, 1854 syllables) and a test set (13 sentences, 531 syllables). Observing the RMSE in the test set an overadaptation to the training data was avoided. Although the network has a fairly simple structure, it is apparently suited to predict the duration and the f0 model points, respectively. The accuracy of prediction mainly depends on the predictive power of the 17 selected input parameters.

The resulting RMSE of the trained network observed in the test set are as follows - *syldur*: 47.1 ms, *paudur*: 56.4 ms and *f0 model point*: 19.4 Hz. Figure 2 shows the measured *f0*-contour of an original Mandarin utterance versus the synthesis-based contours with original f0 (*copy contour*) and network-predicted f0 values.

4. RESULTS AND EVALUATION

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.

- Three sentences were resynthesized with original (natural) phoneme duration, with original syllable duration and with the network-predicted syllable duration, respectively. For the resynthesis experiment the natural f0 structure was kept.

- 9 sentences were synthesized using the TTS system without duration adjustment (duration from the inventory), with original syllable duration and with network-predicted duration, respectively. The intonation structure was predefined by the synthesis inventory. The same sentences were additionally tested with original f0 (*copy contour*) and network-predicted f0 contours.

The listening test consists of total 54 sentences: 9 resynthesis sentences, 36 synthesized sentences, and 9 correspondent original sentences. The overall quality was judged by 20 Mandarin native speakers by an absolute category rating using a scale from 1 to 5 (mean opinion score: MOS).

The evaluation of resynthesis sentences in figure 3 reveals the general impression of duration control with original pitch contours. The original (natural) sentences (*original*) achieved the best rating (MOS=4.91) - only slightly better than the examples basing on network-predicted syllable durations (*nndur*, MOS=4.88). The non-significant MOS differences in resynthesis prove the applicability of the neural network prosodic model on principle.

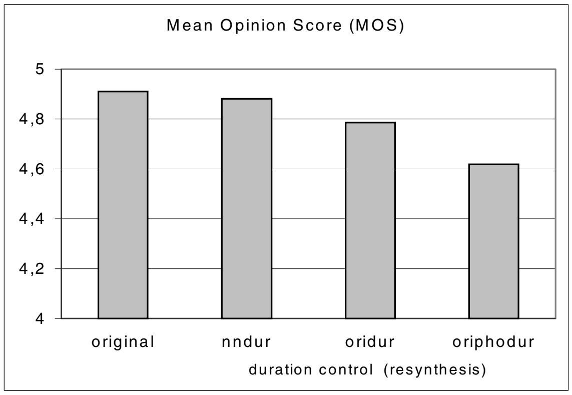

Fig. 3. Evaluation of resynthesis examples generated from manipulated duration values.

Figure 4 reports the results of DRESS text-to-speech synthesis. Without duration and intonation modification (*mono*, prosody taken from inventory), the overall quality can hardly be accepted by the listeners (MOS=1.91). Applying the original syllable duration (*oridur*) or the neural network duration modification (*nndur*) the quality achieves an improvement (MOS=2.25 and MOS=2.48, respectively). The integrated neural network-based duration and intonation control (*nndurf0*) leads to a strong improvement (MOS=3.20) and almost achieves the reference synthesis using original syllable duration and *f0* contour (*oridurf0*, MOS=3.42). The remaining gap between synthesis and original stimuli (MOS=4.91) is caused by PSOLA signal manipulation and by restricted coarticulation modeling.

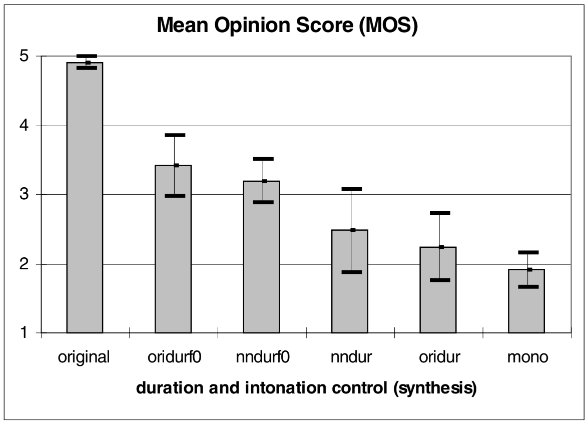

Fig. 4. Evaluation of synthesis examples (modified duration and intonation).

5. CONCLUSION

Using a small but especially designed syllable database and an enhanced linguistic feature set, a neural network approach enables an efficient prosody control in Mandarin Chinese concatenative speech synthesis. The network modification is appropriate and it partly achieves evaluation scores near to natural speech. Perceptual tests of resynthesized stimuli using network-predicted duration values receive an excellent Mean Opinion Score (MOS) comparable to natural speech of about 4.8. Due to the lack of necessary signal manipulation, the tonal coarticulation, which has been successfully predicted by the network, has been inadequately realized in the synthesis. Therefore, the synthesis stimuli (MOS=3.20) is still dissatisfying compared with natural speech. According to the *copy contour* synthesis experiment the evaluation gap between original (natural speech)and synthesis is only partly caused by the neural network prosodic model (δ_{MOS}=0.22).

6. REFERENCES

[1] O. Jokisch, H. Mixdorff, H. Kruschke, and U. Kordon, “Learning the parameters of quantitative prosody models,” *Proc. ICSLP 2000, Beijing*, pp. 645–648, 2000.

[2] H. Mixdorff and O. Jokisch, “Building an integrated prosodic model of german,” *Proc. Eurospeech 2001, Aalborg (Denmark)*, pp. 947–950, 2001.

[3] M. Chu and Y. Feng, “Study on factors influencing durations of syllables in mandarin,” *Proc. Eurospeech 2001, Aalborg (Denmark)*, pp. 927–930, 2001.

[4] Z. Ying and X. Shi, “An rnn-based algprithm to detect prosodic phrase for chinese tts,” *Proc. ICASSP 2001, Salt Lake City*, pp. 809–812, 2001.

[5] S.-H. Hwang et al S.-H. Chen, “An rnn-based prosodic information synthesizer for mandarin text-to-speech,” *IEEE Trans. on Speech and Audio Signal Processing*, 1992.

[6] C. Shih and B. Ao, “Duration study for the bell laboratories mandarin text-to-speech system,” In Santen97 [10].

[7] H. Ding and J. Helbig, “Modeling duration and tonal coarticulation in a mandarin chinese synthesis,” *Proc. ISCSLP 1998, Singapore*, pp. 243–248, 1998.

[8] H. Ding and J. Helbig, “Prosodic alternative units in a mandarin chinese speech synthesizer,” *Proc. ISCSLP 2000, Beijing*, pp. 101–104, 2000.

[9] R. Sproat, Ed., *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, Kluwer Academic Publishers, Dordrecht (NL), 1998.

[10] J.P.H. van Santen, R.W. Sproat, J. Olive, and J. Hirschberg, Eds., *Progress in Speech Synthesis*, Springer, New York, 1997.