TOWARDS A MULTILINGUAL PROSODY MODEL FOR TEXT-TO-SPEECH

Oliver Jokisch, Hongwei Ding, Hans Kruschke
Dresden University of Technology
Laboratory of Acoustics and Speech Communication, 01062 Dresden, Germany
{oliver.jokisch,hongwei.ding,hans.kruschke}@ias.et.tu-dresden.de

ABSTRACT
The generation of prosodic parameters such as F0 contour, duration and intensity still remains an important issue for naturally-sounding text-to-speech (TTS), although recently developed TTS systems have achieved a considerable progress. Several appropriate but language-specific rule-based, statistical or data-driven prosody models have been successfully realized in many systems. The language and parameter dependent models lead to a more complex and inefficient TTS system design. In earlier works the authors proposed a hybrid data-driven and rule-based model, which can adjust different voices or speaking styles by learning and predicting prosodic parameters. The current paper discusses the multilingual model generalization and the design of appropriate prosodic databases. Exemplary, two different languages: German and Mandarin Chinese are examined. Prediction results and perceptual evaluation with respect to F0 contours and duration values are presented. Since the perceptual results of both languages are comparable and satisfying, the model is qualified for the multilingual prosody control. Resynthesis stimuli obtained from modified prosodic parameters partly achieve near-to-natural mean opinion scores (MOS) above 4.0. The introduced hybrid data-driven and rule-based model is comparatively simple and enables a multilingual prosody control in TTS.

1. INTRODUCTION
Prosodic modeling and prosodic modification in TTS are still indispensable. Though both corpus-based synthesis systems as CHATR [1] and systems selecting larger units of speech as BOSS [2] have achieved a considerably sound quality, they still require least prosodic targets or prosodic smoothing algorithms. The goal of improving prosodic models is the generation of naturally-sounding synthesis with modified f0 contours, duration or intensity.

In standard time-domain synthesis systems like DRESS (Dresden Speech Synthesizer [3]) the prosodic modeling belongs to the core tasks. Several appropriate but language-specific rule-based, statistical or data-driven prosody models can be found in literature, but the language and parameter dependent models lead to a fairly complex, inefficient and often inflexible TTS system design.

Data driven algorithms for prosody control enable simple adjustment of prosodic parameters by training and lead to more variable and natural contours. A strictly data-driven method for the parametric multilingual prosody control as introduced with PAPAGENO [4] partly conflicts with the language-specific speech production and the discriminative knowledge representation.

In earlier works the authors proposed a hybrid data-driven and rule-based model (HYDRA) to learn and to predict prosodic parameters. The model can adjust different voices or speaking styles. The current paper discusses the multilingual model generalization and the design of appropriate prosodic databases. Two different languages: German and Mandarin Chinese are examined. Prediction results and perceptual evaluation with respect to F0 contours and duration values are presented. Since the perceptual results of both languages are comparable and satisfying, the model is qualified for the multilingual prosody control. Resynthesis stimuli obtained from modified prosodic parameters partly achieve near-to-natural mean opinion scores (MOS) above 4.0. The introduced hybrid data-driven and rule-based model is comparatively simple and enables a multilingual prosody control in TTS.

2. HYBRID DATA-DRIVEN AND RULE-BASED CONCEPT
Basing on extensive research work - rule based, quantitative prosody models often outperform strictly data driven approaches. Keeping the features of rule based modeling it seems to be necessary to extend the models by parameter-learning and adjusting components. The hybrid data driven and rule based architecture (HYDRA [5]) includes a rule based, quantitative prosody model (the core component), an automatic extractor for the model parameters, a data driven algorithm for adjusting the model parameters, an interface to the according database (training time) and to the TTS text processor, respectively. During the training a linguistic-phonetic feature vector is applied to the input of the data-
driven module while the output is representing the parameter vector (e.g. initial time, amplitude, etc.) at the input of the rule-based module (supervised learning). The HYDRA approach can be considered as rule-based model extended by learning procedure. The data-driven module only acts as a "controller" enabling the adjustment of mean values, f0 ranges, etc. for new languages, speakers or speaking styles.

Following the HYDRA approach figure 1 demonstrates an example, in which the known Fujisaki control parameter set \( (T_1, T_2, A_a, T_0, A_p) \) is predicted by a feed-forward neural network (FFNN).

3. MULTILINGUAL PROSODIC DATABASES

3.1. German database

The German database used in this study is part of a German speech corpus compiled by the Institute of Natural Language Processing at the University of Stuttgart [6]. It consists of 72 broadcasting news stories read by a male speaker. The total recording time includes 48 minutes of speech containing 13151 syllables. The corpus contains boundary labels for phones, syllables and words as well as ToBI-labels.

3.2. Chinese database

There are many factors controlling prosody, different speakers can still introduce immense variations to the complexity. To train the database for the purpose of modeling the prosody in synthesis, the database were uttered by the same Mandarin male speaker. The source of Chinese database consists of two subsets: phonetically-labeled synthesis inventory and prosodically-labeled natural text. Total 3049 syllables of the inventory were segmented from carrier sentences. Most syllables carry phrase stress and are followed by neutral tones, so that inherent tonal contours are least disturbed from neighboring syllables. These characteristics proved to be important for a time domain synthesis. The natural text database contains 63 sentence of 15 to 86 syllables per sentence (total 2385 syllables), which were selected from one hour reading of newspaper articles for broad coverage of prosodic factors.

Investigation of the important prosodic factors are focused on four aspects: phoneme identity, lexical tones, stress, and position. To cover these factors, the database labeling was performed in three tiers: prominence, location, acoustic segment and its tone. All phonetic and prosodic labeling were processed manually to minimize errors, which can be effected by automatic segmentation. After several statistic procedures, a data matrix was resulted. Each syllable or pause is attached with its contextual factors, duration and f0 values. They include syllable type and tone; the distance from the beginning and end of the word, phrase and utterance; the previous tone and next tone; the prominence feature and duration value, f0 was measured at the left margin, in the center and at the right margin of the syllable. Great effort was contributed in the examination of the prosodic factors with a great deal of manual performance to provide reliable information for neural network, which is essential to the accuracy of prosodic prediction.

The analysis results of both German and Chinese databases were arranged in a syllable-oriented table for the following model training.

4. MODELS AND TRAINING

To match multilingual requirements with a simple uniform model the authors act on a few assumptions:

- The syllable frame is the appropriate processing level.
- Approximately, 20 independent linguistic and phonetic factors (input) affect the prediction of the prosodic model parameters.
- These factors act simultaneously. Time recurrence is not necessary (For each step context information is available).
- 3-5 model parameters (output) per prosodic parameter are sufficient.

4.1. Integrated model of German

Mixdorff [7] introduced an Integrated Model of German (IGM) predicting syllable duration along with syllable-aligned Fujisaki control parameters (f0) and using the hybrid data-driven and rule-based processing scheme discussed in section 2. IGM results indicate synergy effects in case of an integrated prosodic processing and prove the importance of an extensive linguistic analysis. A FFNN is predicting a total of 8 model parameters per syllable from a set of 24 input features (figure 2).
4.2. Syllable oriented model of Chinese

Tone language, such as Chinese, has lexical tones in syllable to distinguish meanings. The f0 prediction for the syllabic microintonation is essential to the intelligibility. The resulting f0 movement is a superposition of the segmental tone contours and suprasegmental phrase declination, which is illustrated in figure 3 [8]. Due to sentence lowering effects and tonal coarticulation, segmental tone height and shape are affected. The amount of effects can be captured through f0 determination at several key points in a syllable. The left and right margin were assumed important for tonal coarticulation, the syllable center was regarded as reference point for tone height. The FFNN will train the tones (at these points: $f_{\text{left}}$, $f_{\text{cent}}$, $f_{\text{right}}$) as well as the duration values of syllables and pauses with their contextual features from the labeled database. The total number of output parameters is 5.

The Chinese database (2385 syllables) was also subdivided into a training set (1854 syllables) and a test set (531 syllables).

5. RESULTS AND EVALUATION

The resulting RMSE (between original and network-based parameters) observed in the test data of German and Chinese are very similar (RMSE=0.046 s versus RMSE=0.047 s for the syllable duration, RMSE=0.069 s versus RMSE=0.056 s for the pause duration). The RMSE of the network-predicted and model processed f0 values is less for German (RMSE=17.8 Hz) than for Chinese (RMSE=24.8 Hz). A series of perception experiments for evaluating the quality of the German model (IGM) were designed (description in [9]). By applying network-based modifications in duration and f0 contour, test stimuli using PSOLA resynthesis (PRAAT software) and MBROLA diphone synthesis were created. 21 German native speakers judged the overall quality of 192 different stimuli with an inverse Mean Opinion Score (MOS) from 5 (very bad) to 1 (very good). The Chinese listening test (only modified syllable duration applied) consisted of total 54 sentences using PSOLA resynthesis and synthesis (from DRESS). The overall quality was judged by 20 Mandarin native speakers on a scale from 1 to 5 (MOS).

Fig. 2. FFNN predicts 8 model parameters from a set of 24 input features.

Fig. 3. Principle of the superposition of syllable inherent tone contours and the sentence intonation pattern [8].

Fig. 4. Evaluation of German (resynthesis and synthesis).
Figure 4 shows the results of German. The original stimuli achieved the best rating (original, MOS=4.51). Resynthesis examples by modifying either duration or f0 contour (resyn_origf0, resyn_origf0-resynsyl) also achieved good MOS of 4.08 and 3.76, respectively. The manipulation of both prosodic parameters in resynthesis (resyn_origf0-nnsyl) still leads to acceptable results (MOS=3.36), whereby the corresponding stimuli in synthesis (syn_origf0-nnsyl) is degraded by the poor acoustic quality (MOS=1.75).

The evaluation of resynthesis sentences in figure 5 reveals the general impression of duration control with original pitch contours. The network-based sentences (resyn_nnsyl) achieved the best rating (MOS=4.88), surprisingly even a slightly better than the original (natural) ones (MOS=4.83) and also better than the resynthesis sentences of the original syllable duration (resyn_origsyl, MOS=4.79).

Figure 6 reports the results of duration control without intonation modification (synthesis). Without duration modification (syn_monoton), the overall quality can hardly be accepted by the listeners (MOS=1.7). Applying the original syllable duration (syn_origsyl), the quality achieves a little improvement (MOS=2.24). The neural network duration modification (syn_nnsyl) leads to a further improvement (MOS=2.48), since the network averaged the strong variations of the original duration. The intonation control was not tested in the synthesis experiment. The missing intonation modification is mainly responsible for the large gap between synthesis and original stimuli (MOS=4.77).

6. CONCLUSION

The introduced hybrid data-driven and rule-based model is comparatively simple and easily adaptable. The perceptual results in both German and Chinese resynthesis stimuli are convincing and implicate that this model is appropriate for a multilingual use, provided that a TTS system can generate the necessary linguistic information from the text processing. The successful application of this prosody model in synthesis demands an improved signal manipulation.

7. REFERENCES