An Adaptable Acoustic Architecture in a Multilingual TTS System

Caglayan Erdem[✳], Janez Stergar[✦], Bogomir Horvat[✦]

(✳) Siemens Corporate Technology, Dept. CTIC 5, 81730 Munich, Germany
(✦) University of Maribor, Faculty of EE and Comp. Science Maribor, Slovenia
caglayan.erdem@bmw.de, (janez.stergar, bogo.horvat)@uni-mb.si

Abstract

In this paper an adaptable acoustical architecture in a multilingual TTS system is presented. The whole architecture is designed to be a data-driven system. Modules comprising text preprocessing, grapheme-to-phoneme conversion, lexical stress detection, OOV-handling, symbolic prosody prediction, acoustic prosody prediction and unit selection with concatenation use machine learning techniques especially neural networks (NN) or language independent routines. The adaptable and scaleable architecture of the acoustic prosody generation module is built up by four sub-modules. While duration control uses a NN designed on the modified causal error correction architecture (CRCECNN), f0-generation utilizes a MLP NN. Within both NN modeling a partially Weight Decay (p-WD) method is applied to optimize each input vector dimension of the NNs. The p-WD method helps to select one of the highly correlated features in contrast to standard weight decay; hence through its penalty function we achieved a minimized input feature set. By the use of the third sub-module, which reuses the predictions of the optimized NNs, a hybrid architecture is established, as unit selection based on syllable prosody parameter criterions combines prosody selection with unit selection. Handling with a limited database makes a post processing unit necessary. We'll emphasize the problem of finding optimal speech segments and an approach of segment selection using a global parameterized non-linear suitability function in combination with a modified multi-level Viterbi search algorithm. Preliminary acoustic ratings of the adapted TTS system to Slovenian language will be introduced.

1. Introduction

Automatic learning techniques offer a solution in adapting a TTS system to a new language, voice or a new application. They allow automatic extraction of specific features (e.g. non-uniform unit selection, prosodic regularities extraction) from an appropriate database of natural speech. Such techniques depend on the construction of a large preprocessed corpora (properly segmented, labeled with appropriate symbolic prosody labels, etc.). The preprocessing and labeling can be performed either automatically or by hand. While automatic labeling can be less accurate than hand labeling, the latter is very time consuming (and in some tasks also inconsistent). However in some processes, such as segmentation to non-uniform units, which are crucial for concatenative TTS synthesizers and verification of automatically labeled data, expert guided procedures cannot be avoided. On the other hand, many adaptation tasks can be realized automatically or semi-automatically. We will introduce methods used to adapt the acoustical part of a multilingual TTS system to Slovenian language. Therefore in the following sections the module for acoustic modeling will be disseminated.

We'll introduce the neural network (NN) structures suitable for adaptation to a new language without language expertise. The implementation of a modified weight decay method to overcome the known overfitting problem in the process of NN learning which considerably reduces the difficulties of data-driven adaptation will be applied. The proposed method is based on soft input pruning, fading out the unimportant inputs. The so-called p-WD was implemented in the f0-generation module. Also the module for duration modeling of concatenation segments will be presented. The modified causal retro-causal NN will be introduced and the removing training connection procedure explained. We'll also introduce the unit selection module, which is based on a modified multilevel Viterbi-search algorithm. Because of the limited database used for adaptation also the implemented post processing method will be introduced.

2. The database

The database (corpus) used for prominence modeling consists of app. 1200 sentences in the Slovenian language (approx. three hours of speech). The selection of the text was emphasized for the broad coverage of sentences in the Slovenian language with the main concern towards the best coverage of concatenation segments.

The audio database recordings were created in a studio environment with a male speaker reading aloud isolated sentences in the Slovenian language (44.1 kHz, 16 bit).

The whole corpus was designed using a selection of clauses from a 31 million word corpus in the Slovenian language. The major parts of the clauses covered daily-published news and Slovenian literature; the minority consisted of clauses taken from Slovenian poetry.

First, sentences not shorter then 15 and not longer than 25 words were preselected from the major corpus. Then, four different text corpora were generated and analyzed statistically (approximately 5000 sentences per corpus). The selection of sentences for the final corpus (database) was based on a two-stage process. In the first stage an analysis based on statistical criteria was performed. In the second stage the final text was chosen based on the results of the first stage. In the final database 1200 sentences remained.

The phonetic transcription was managed using a two-step conversion module. The first step is realized with a rule-based algorithm. The second step was designed with a data-driven approach (NNs were used) [2].

Pronunciation was derived from the IPA Alphabet. In order to represent the IPA symbols in ASCII characters the SAMPA format was widely used. In our grapheme-to-phoneme conversion module the SAMPA phonetic transcription symbols for the Slovenian language were used [9].

The text corpus was hand-labeled using 13 different classes of part-of-speech tags (POS). All tags were combined in an environment where tracking and correcting tags was simplified for the labeler [3].

The spoken corpus was phonetically transcribed using HTK. Entities "sil" and "sp" (short pause) respectively, denoting the silence before and after a sentence and between words were determined with a one-state HMM and all phonemes with three-state HMM in the HTK environment.

3. The acoustic modeling architecture

In the following the four modules constituting the adaptable acoustical architecture in the used multilingual TTS system PAPAGENO will be briefly explained (*Figure 1*).

3.1. Duration modeling

The NN depicted in *Figure 2*, utilizes shared weights and finite unfolding. The coupling of both information flows is realized by only one output cluster z_t instead of the coupling at each time step within CRCECNN [4]. By coupling those information flows within the present time step this new architecture does not contain fix-point recurrent loops, which might cause instabilities during training. In the following this architecture will be used for further adaptations applying structural switching.

During training all segmental durations modeled as observations are known. But within the application there are no observations available for $i \geq 0$, because they are not predicted yet. For $i < 0$ predictions of the NN are re-utilized as observations. Because of this mismatch between training and application the retro-causal information flow has to be treated in a specific way. We approached to the problem with two modified architectures. The approach using the asymmetric P-CRCECNN combined with the procedure of removing con-
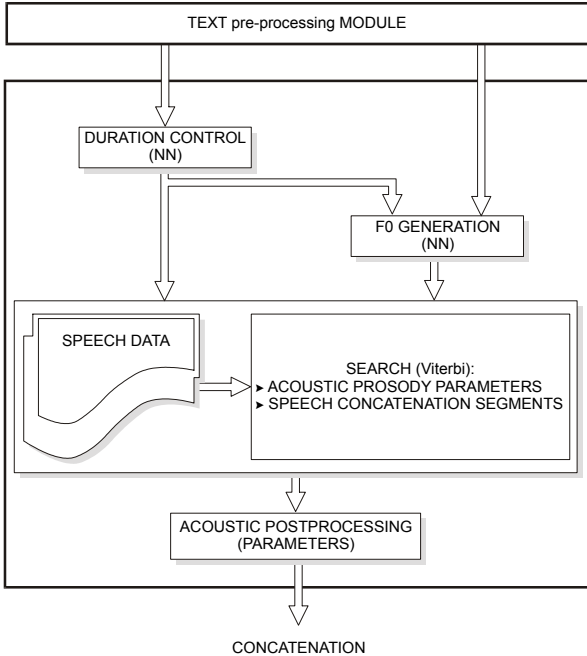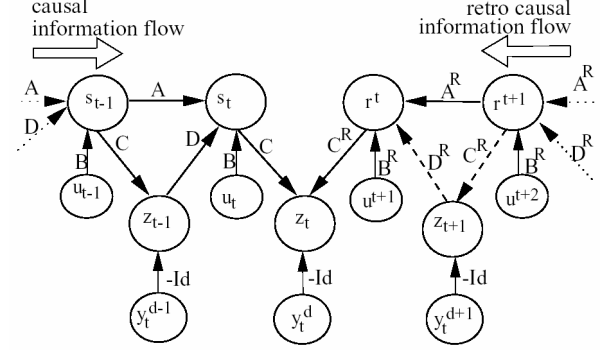

Figure 2: *The architecture of P-CRCECNN. Indices R denote the Retro-Causal-Modeling-Path.*

nections after training showed the best result [4]. For training the same architecture is used as depicted in *Figure 2*. The dotted connections C_R and D_R are trained. So the basic architecture remains the same during training, but within the application connections C_R and D_R are removed. The resulting architecture is then a finite unfolding in time without the error correction principle for the retro-causal information flow during application.

In the following the application of asymmetric P-CRCECNN's within the segmental duration control unit of our acoustic prosody NN module is presented. The used data-driven methods are applied to recordings of approx. three hours of a Slovenian database as already described (section 2.1 The corpus) [8]. The used database (cf. section 2) is the same as applied within the f0-generation task. The f0-generation task utilizes patterns organized on syllable level – within this task, patterns are organized on triphone level.

The following information (extracted from the database) is presented to the NN input in a context of seven phonemes to the left and right:

- Phonetic information: with one-out-of-n coding the phoneme index is presented here. A phoneme-set of 45 phonemes is used. Additionally the four phoneme classes (vowel, fricative, nasal, liquid, and plosive) are presented .
- Positional info. discrete info.: denotes whether the according syllable is an initial, medial or final one within the phrase and the word. Continuous info. is given by the relative syllable position within a sentnce and phrase.
- Stress info.: flags denoting the stress type of the according syllable are coded here. Four flags present word level stress. Sentence level stress consists of two stress marks (predicted in a separate module).
- Ling. cat.: a one-out-of-n (set of 13 categories) coded linguistic category part-of-speech (POS) denotes the category type of the according word.

All listed input categories are presented at each time step of the unfolding clusters denoted by u_{t+1} (*Figure 2*). The according output vectors are modeled as observations and are presented at each time step in the clusters denoted by y^d_{t+1}. Target values for the NN are normalized to ensure an optimized signal-flow during training of the NN due to tanh activation function within the causal and retro-causal state clusters. A first normalization of segmental duration is obtained by the mean and standard deviation value from the used triphone classes. A second normalization was necessary to ensure an

Figure 1: *The adaptive acoustical architecture of the multilingual TTS system.*

optimized signal flow during training of the NN. The mean and standard deviation were derived from the first normalized segmental durations.

In the adaptation procedure the patterns for training (80%) and testing (20%) were separated. A validation set of (20%) was selected randomly from the training set. For evaluation the trained NN were used to predict segmental durations of sentences that were in the test set.

3.2. F0-contour generation

The p-WD [5] has been applied to the f0-contour generation module. The utilized NN has to map input parameters to an appropriate f0-contour. Regarding the syllable the mapping is performed to four f0-contour parameters (*Figure 3*). The solid line depicts a f0-contour on the syllable level. These contours are parameterized (dashed line) by four values: f0-maximum ($p1 = F0_{MAX}$), f0-maximum position ($p2 = F0_{MAX\_POS}$), f0 at syllable start ($p3= F0_{START}$), and f0 at syllable end ($p4= F0_{STOP}$). For the contour parameterization a maximum based description is used, which mainly defines that f0-contours on syllable level for non-tonal languages can be described by a rising on the first part and a falling on the second part of the syllable [6].

The mentioned parameters p1, p2, p3, and p4 are the outputs $y = \{p1, p2, p3, p4\}$ of the NN respectively. Hence the dimension of the output cluster m = 4.

F0-contours are known to be influenced by long-term features (the sentence type), breath and local stress intention. The input parameters must contain information concerning local and global characteristics (symbolic prosody tags). For a good mapping it is also important to provide contextual information of the syllable. Due to computation reasons the context window length was chosen to be seven to the left (past) and seven to the right (future) of the syllable with the exception of the linguistic categories. The following input features are presented to the NN to solve this problem on the syllable level for each context unit:

- Phonetic information: The phonetic structure of a syllable to be processed is coded here. The vowel is presented as a one out-of-n coded input using the Slovenian SAMPA phoneme set. Neighboring phonemes of the vowel are given in four classes (plosive, fricatives, nasal and liquids) and also as a one-out-of-n coded input in a symmetric context window of four phonemes.
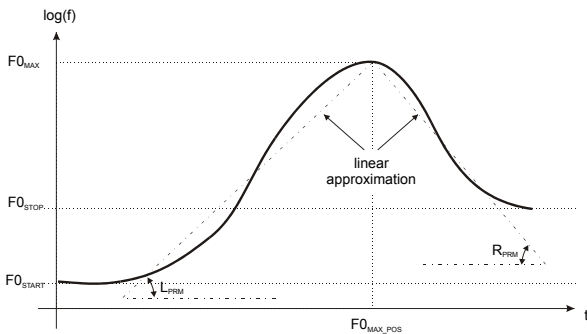- Positional info.: Continuous pos. information gives time

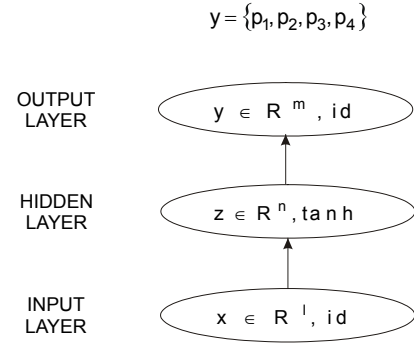Figure 3: *Maximum based parameterization of f0-contours.*

Figure 4: *NN structure for f0-contour modeling.*

distances of the syllable and its vowel. Discrete info. denotes whether this syllable is an initial, medial or final one within the sentence, the phrase, and the word.
- Stress info.: Flags denote the stress type of a syllable within the word and the phrase.
- Ling. cat.: The used linguistic category set consists of 13 tags, which are one-out-of-n coded and presented in a context of 3 to both sides.

Hence by this input constellation the input dimension of x is in the range between 500 and 600 (e.g. 560). The p-WD technique was applied to recordings of three hours of a Slovenian news speaker reading isolated sentences from a large corpus as described in the foregoing section (cf. section 2 The database).

The patterns for training (80%) and testing (20%) were separated. A validation set of (20%) was selected randomly from the training set. The introduced parameter p in the weight decay penalty term of p-WD was optimized by experiments with varying parameters p. This tuned NN module was then used to analyze the inputs and optimize the input feature selection.

3.3. The unit-selection module

The unit selection module within the introduced multilingual TTS system uses a robust unit selection method based on syllable prosody parameters optimization (RUSSPP) [7].

First isolated NN predictions of f0-contours and segmental durations are performed and then these parameters are re-utilized for a search in speech data (corpus) for best fitting of speech segments and acoustic prosody parameters. This search is realized by using a modified Viterbi-algorithm that operates on syllable level using syllable level optimality criterions. But it explicitly allows higher and lower levels of speech segments in the path search procedure.

3.4. The module for post-processing

Dealing with limited speech data (segments) for synthesis makes signal processing on speech elements at concatenation points unavoidable. Therefore we used simple but efficient post-processing on the selected segments prosody parameters. This new method was already applied and tested within the TTS system PAPAGENO for German (male news speaker) [7]. It could be shown that it improves the quality of the used prosody generation module and of the selection process.

It was observed that the used NNs are giving good prosody modeling results within macro prosody. Therefore the

general idea of this post-processing is a realignment of the obtained f0-contours according to the run of the f0-maxima of the triangles as depicted in Figure 3.

After the post-processing, the f0-contours are then realized by modifying the speech-elements using a PSOLA like algorithm for speech synthesis.

4. Experiments

The acoustical results of our adapted multilingual TTS system were presented to a group of 20 non-expert listeners. We generated an inventory of 216 test sentences not used for the training or validation process.

We additionally implemented a module for symbolic prosody tags prediction into the architecture of acoustic modeling. The selective prediction method used essentially contributed to naturalness of the synthesized speech without influencing the intelligibility of synthesized sentences.

The test performed during 3-hour session (approx.) indicates that the adaptable acoustic architecture used in the approach of adapting a multilingual TTS to Slovenian language (based on adaptable NN architectures) is suitable and promising for multilingual speech synthesis. During the listening test each sentence was estimated with marks from 1-5, with 5 denoting the acoustically most pleasant sentence and 1 reserved for unacceptable ones. The average ratings (the variances and ratings for each test person are presented in *Figure 5*) were good-very good (3,276995).

5. Conclusion

In the foregoing sections we introduced an adaptive and scaleable architecture for acoustic prosody generation in a multilingual TTS system. By using only the NN predictions without the prosody selection and post-processing the architecture might be scaled down.

In the introduction we briefly explained the design of a suitable database used for adaptation of all modules in the TTS system. In more detail we explained the designed adaptable acoustical architecture combined of four modules. The first module introduced was the duration control NN module. We emphasized its basic structure with the new p-WD method applied. By its penalty function we achieved a minimized input feature set. The NN duration control module introduced uses the modified causal retro-causal error correction architecture (P-CRCECNN).
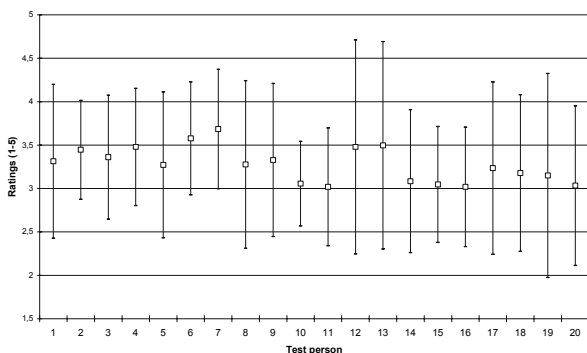

Figure 5: *Values and variations of values in the acoustical test ratings per test-person.*

Considering time-dependency-structures is crucial for duration control task. By this architecture this a priori knowledge is explicitly modeled. Without the finite-unfolding technique it is not possible to consider the time-dependent-information-structures.

The performed acoustical experiments confirmed the suitability of the P-CRCECNN as an adaptable architecture. We observed the acoustic suitability for German as well as for Slovenian solving first the duration control task and afterwards the f0-generation task which helps avoiding problems caused by the strong influence of duration control on f0-generation.

We also introduced the module for unit selection using a new selection method (RUSSP) which is based on prosody parameters optimization. The problem of finding optimal speech segments was also emphasized.

We proposed an approach of segment selection using a global parameterized non-linear suitability function in combination with a modified multi-level Viterbi search algorithm. The preliminary acoustical tests confirm the suitability of the designed architecture for adaptation to a new language and encourage the use of the selective symbolic prosody tags for a more subtle prominence modeling.

6. References

[1] Stergar, J.; Horvat, B., 2003. An environment for word prominence classification in Slovenian language. *In proceedings of the ICPhS03*. Barcelona, Spain, 2087-2090.

[2] Rojc, M.; Kačič, Z., 2000. Design of Optimal Slovenian Speech Corpus for use in the concatenative Speech Synthesis System, *LREC 00*. Athens, Greece, 321-325.

[3] Stergar, J.; Hozjan, V.; Horvat, B., 2003. Labeling of Symbolic Prosody Breaks for the Slovenian Language. *International Journal of Speech Technology*. Vol. 6, No. 3, 289-300.

[4] Erdem, C.; Zimmermann H. G., 2002a. Segmental duration control by time delay neural networks with asymmetric causal and retro-causal information flows. In *European Symposium on Artificial Neural Networks (ESANN)*. Brugges, Belgium.

[5] Erdem, C.; Zimmermann, H.G., 2002b. A Data-driven method for input feature selection within neural prosody generation. In P*roceedings of ICASSP 02*. Orlando, Florida.

[6] Heuft, B.; Portele, T.; Höfer F.; Krämer J.; Meyer, H.; Rauth, M.; Sonntag, G., 1995. Parametric Description of F0-Contours in a Prosodic Database. In *Proceedings of the ICPHS 95*. Vol. 2, 378-381.

[7] Erdem, C.; Beck, F.; Hirschfeld, D.; Hoege, H.; Hoffman R., 2002c. Robust unit selection based on syllable prosody parameters. *IEEE 2002 Workshop on Speech Synthesis*. Santa Monica, California USA.

[8] Stergar, J.; Erdem, C.; Rojc, M.; Kačič, Z.; Horvat, B., 2003. Prosody adaptation and modeling in a multilingual TTS system. Submitted to the *Journal of Machine Learning (Special Issue on Learning in Speech and Language Technologies)*.

[9] Kačič, Z.; Zemljak, M., 1999. SAMPA - computer readable phonetic alphabet. *The WEB portal of Department of Phonetics and Linguistics*. University College London (http://www.phon.ucl.ac.uk).