TEXT–TO–SPEECH SYSTEM MODIFICATION BY SOFT COMPUTING METHODS

Jan Čermák

Department of Telecommunications, Brno University of Technology, Purkyňova 118, 612 00 Brno, Czech Republic, phone: +420 541 149 217, e-mail: cermak4@kn.vutbr.cz

Machine–human communication is becoming more and more popular by the public and also by commercial sector. One of the possible machine–human communications is a transformation of a text to speech. Systems performing this transformation, usually with unrestricted dictionary, are called Text-to-speech (TTS) systems. Nowadays, TTS systems are utilized in web browsers, email reading application, telephone services etc. But although the TTS systems usage is very wide, they are still not often used. This is mainly caused by insufficient prosody modeling of synthesized utterance. Prosody modeling is a language dependent complex task, which has not been sufficiently solved yet. Scientists are trying to find new approaches to improve prosody modeling. One of the possible approaches is introduced in this paper.

Keywords: TTS system, fuzzy logic, soft computing, prosody

1. INTRODUCTION

TTS systems can be basically divided into two main functional blocks – linguistic analysis and speech synthesis. See figure 1 for linguistic analysis block scheme [5]. Prosody modeling follows after input text parsing and phonetic transcription. The output of prosody modeling block is represented by prosody parameters – fundamental frequency, sound duration and sound intensity.

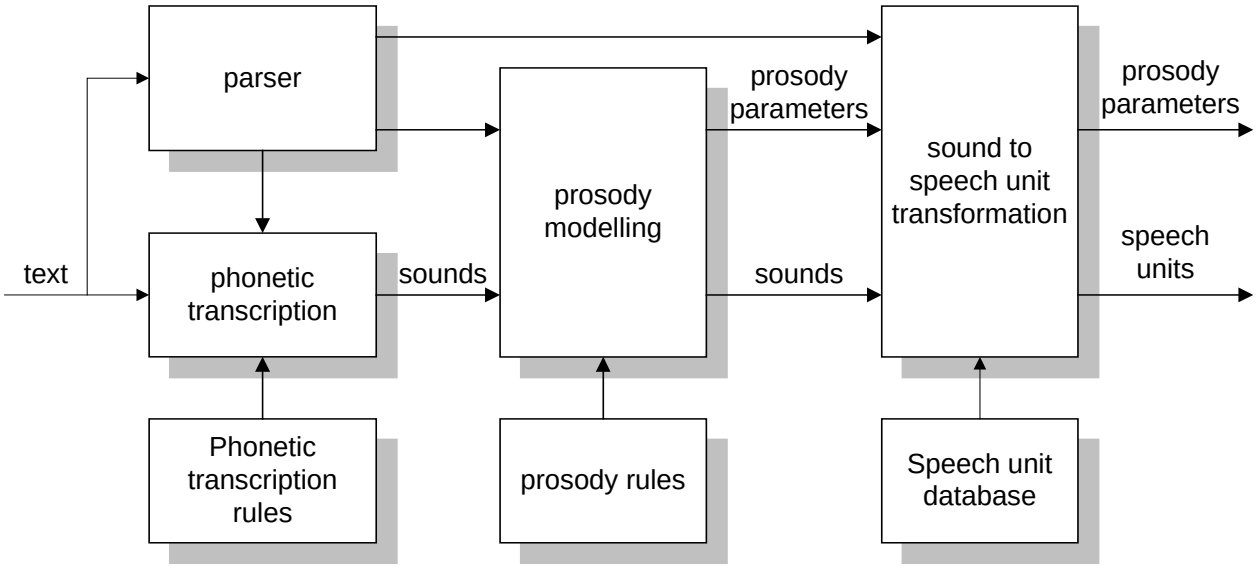

Figure 1. Linguistic analysis part of TTS system.

The fundamental frequency f_0 is perceptually the most important parameter and can be easier modeled then the other parameters. There are more approaches for prosody modeling but usually it is done by rules that are crisp, deterministic and

precise. However the speech prosody is vague and can be hardly described by equations and crisp numbers. In our approach we use one of soft computing methods – fuzzy logic for prosody modeling.

The sound-to-speech unit transformation block converts the sounds into the system speech units – usually diphones or triphones. Prosody parameters and speech units form input of speech synthesis block, that can be based on different methods e.g. PSOLA, LPC.

2. FUZZY LOGIC

The concept of Fuzzy Logic (FL) was introduced in 1965 by professor Lotfi Zadeh at the University of California in Berkley [1], and was presented as a possible way of processing data by allowing partial set membership rather than crisp set membership or non-membership. FL concept was then developed in FL theory, which is used in many scientific disciplines (controlling, economy etc.) now. The best–known and the oldest goal of FL theory is modeling of uncertainty.

Fuzzy system (FS) is a complex system whose variables (at least some of them) are not crisp numbers but fuzzy numbers that are defined on a given universe. The main aim of FS is to control a complex process. The control is based on a set of rules that are representing knowledge of the controlled process behavior. FS input is a vector of crisp numbers $\boldsymbol{X}$. The crisp numbers, usually defined on different universes, are representing the state of the controlled process. The system output is a vector of crisp numbers $\boldsymbol{Y}$ that regulate the controlled process. Basically, there are two different types of fuzzy systems (controllers) – Mamdani and Sugeno. We used the Sugeno based controller – a modification of Mamdani controller. Fuzzy rules of Sugeno based controller have fuzzy antecedents and crisp consequent contrary to Mamdani system with both antecedent and consequents described by fuzzy variables

Fuzzy controller design involves decisions about a number of important parameters that should be determined before the actual control starts. A sophisticated design of complex Sugeno controller is to determine the number of rules and membership functions by a clustering method and consequent functions by methods solving the least squares problem. However this design is just approximative and needs further parameter adjustment by adaptation process e.g. backpropagation, least mean square method [2]. Although backpropagation method was originally developed for neural networks it can be also used in FS because FS can be modeled as a feedforward neural network, see figure 2 for Sugeno type controller ANFIS (Adaptive Neuro Fuzzy Inference System).

The function of each layer shown in figure 2 is following [4]:

- Layer 1 is the input layer. This layer only transmits the input values to the next layer. No computations are done in this layer.
- Layer 2 is the fuzzification layer where each node corresponds to one membership function. The degree of membership is computed here.

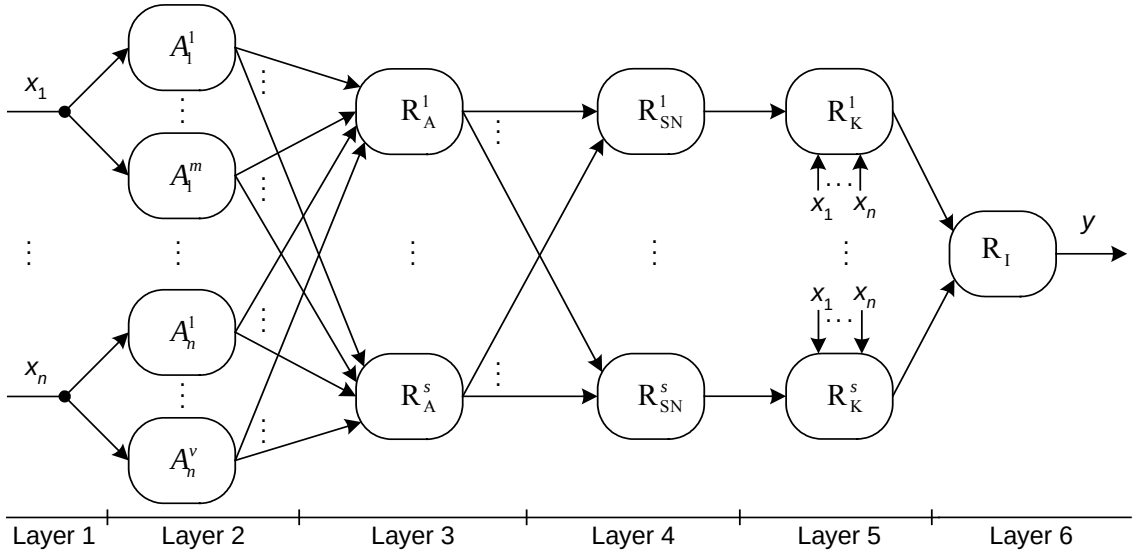

Figure 2. Sugeno controller represented by feedforward neural network.

- Layer 3 is the rule antecedent layer. The antecedent is evaluated in this layer. The output of this layer represents the strength of the corresponding fuzzy rule. Each node corresponds to one rule.
- Layer 4 is the rule strength normalization layer. Every node in this layer calculates the ratio of its rule strength to the sum of all rules strength.
- Layer 5 is the rule consequent layer. Every node in this layer counts the consequent as the product of layer 4 outputs and the consequent function.
- Layer 6 is the rule inference layer. The single node in this layer computes the system output as the summation of all incoming signals.

Sugeno system is implementing the following equation

$$f(\boldsymbol{x}) = \frac{\sum_{l=1}^{s} y^l \prod_{i=1}^{n} A_i^l(x_i)}{\sum_{l=1}^{s} \prod_{i=1}^{n} A_i^l(x_i)}, \qquad (1)$$

where s is the number of the rules, n is the number of input variables, y^l is the consequent function. Membership functions $A_i^l(x_i)$ are the fuzzified input vector variables $\boldsymbol{x} = [x_1,...,x_n]$.

3. PROSODY MODELING

Prosody is usually modeled by rules, Fujisaki model etc. Although alternative approaches were already introduced. The experiments on prosody modeling by neural networks were successfully done in [3]. We used the similar approach because the FS can be implemented as a neural network. Such a system has the properties of both systems – learning ability and knowledge representation. It means that the trained

system is not represented by "black box" as in case of neural network, but the system properties are represented by knowledge.

We decided to model prosody by ANFIS, because it is easy to use and gives sufficient summary of the FS possibilities, advantages and disadvantages. In order to design ANFIS, it is necessary to prepare a training corpus. We used training corpus consisting of 12 Czech declarative sentences. Each sentence was converted into the seven vectors giving the following parameters for each sound:

- sound position in the syllable,
- sound count in the syllable,
- syllable position in the word,
- syllable count in the word,
- word position in the sentence,
- word count in the sentence,
- fundamental frequency.

Fundamental frequency formed the output vector and the other parameters were used in input vector. Fundamental frequency was interpolated for non-voiced sounds in order to have continuous f_0 curve even if non-voiced sounds do not have f_0. System rules and membership functions were estimated by subtractive clustering method and the consequent functions were determined by linear least squares method. Then the hybrid method (least mean square and backpropagation) was used for system training. The training-stopping criterion was related to the number of epochs – 1200 training epochs were always sufficient for our approach.

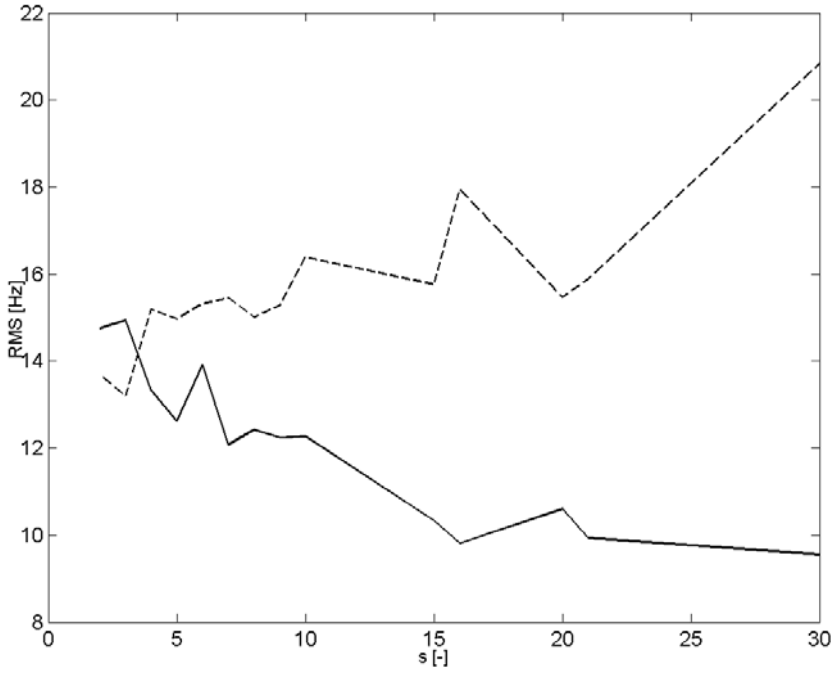

Figure 3. Training results – *s* is the number of input fuzzy sets, training data RMS – solid line, testing data RMS – dashed line.

The quality of generated fundamental frequency was evaluated by root mean square (RMS) value. The relation between the number of fuzzy sets on each input universe *s* and the RMS is shown in figure 3. The solid line represents the trained FS RMS relating to the training corpus (RMS between generated f_0 and original f_0 used for FS training). The dashed line is the RMS relating to the testing corpus consisting of two declarative sentences. It can be seen in the figure 3, that the RMS of the training corpus is decreasing because the fundamental frequency can be more precisely described by higher number of fuzzy sets. Contrary to the training corpus, the RMS of testing corpus has increasing trend. The FS with high number of fuzzy sets starts to be too precise and is not able to model sentences not included in training corpus. This kind of FS even generates extremely high or extremely low f_0 causing high RMS for some input combinations not presented in training corpus. In order to limit the RMS we used threshold on 50Hz and 170Hz. But even with the threshold limits the RMS was not reduced under the RMS caused by FS with low number of fuzzy sets.

The f_0 generated by FS with 5 membership functions and 79 nodes and the original f_0 is presented on Czech sentence "The file is saved – Soubor je uložen" in figure 4. Although the error, which can be seen from the figure 4, is not small, the shape of the generated curve is similar to the original one. The RMS error in our approach is close to the RMS error presented in [3], but we used smaller training corpus and testing corpus. The RMS error is often used for evaluation because it can be counted automatically, however, it does not concern human perception of the generated voice. It would be also necessary to listen to the generated voice for final evaluation.

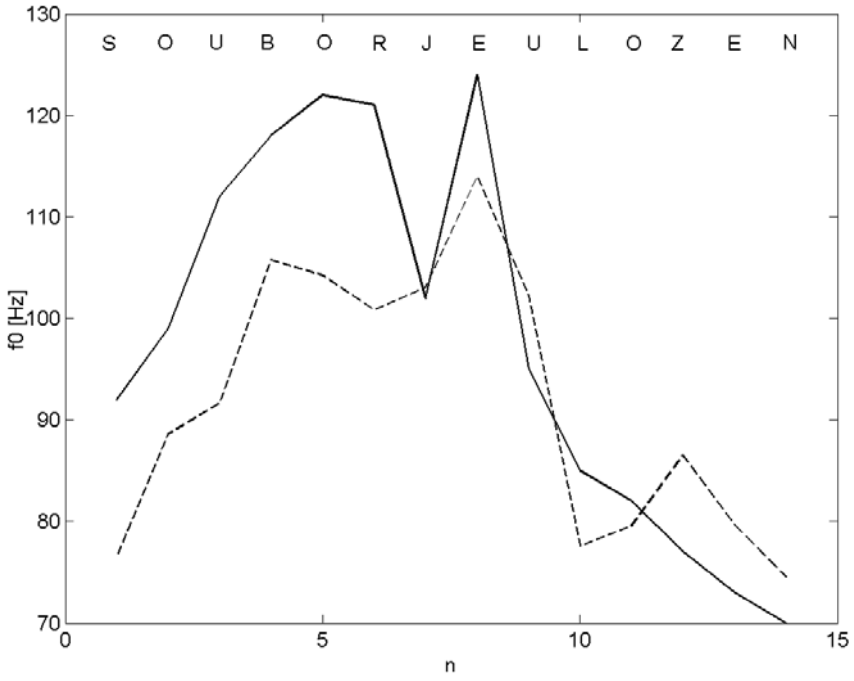

Figure 4. Original fundamental frequency –solid line, modeled fundamental frequency – dashed line.

4. CONCLUSION

The aim of this paper was to introduce the possible approach for prosody modeling in TTS systems. In last chapter, we presented the FS usage for prosody modeling and the achieved results but it would be advantageous to use more extensive training and testing corpus for FS evaluation. FS has similar results to neural network, but the training process is less time consuming and the trained system represents the knowledge of the controlled process. However the knowledge presented by ANFIS system is difficult to use for prosody analysis. FS with less membership function model the sentences not included in training corpus with smaller error and prevent generating extreme fundamental frequency values. In order to avoid the described problems the Mamdani type of FC could be used.

5. REFERENCES

[1] Zimmermann H., Fuzzy *Set Theory – and its applications. Dordrecht:* Kluwer Academic Publishers, 2001. 514 p. ISBN 0 7923 7435 5.

[2] Jura P., *Introduction to Fuzzy logic for controlling and modeling.* (in Czech) Brno: VUTIUM, 2003. 132 p. ISBN 80-214-2261-0.

[3] Adámek, J., Horák P., Sobe D., *Using neural network to model prosody in Czech text-to-speech synthesis.* 13th Czech-German Workshop. Prague, 2004. pp. 59-68.

[4] Abraham A., *Cerebral quotient of neuro-fuzzy systems.* Computing and Internet for the majority of the world, 8th online issue, 2001.

[5] Horák P., *Modeling of Czech suprasegmental features by linear prediction.* (in Czech) Dissertation thesis, ČVUT, 2002.

[6] Ruan, D., *Intelligent Hybrid Systems: Fuzzy Logic, Neural Networks, and Genetic Algorithms,* ISBN: 0792399994.

[7] Narayanan S., Alwan A., *Text to Speech Synthesis.* Prentice Hall PTR, 2004. ISBN 012135661X.