euronoise

Modelling acoustic parameters of prosody for read and acted-speech synthesis

Milan Rusko[a], Marián Trnka[a], Sakhia Darjaa[a], Richard Kováč[a] and Juraj Hamar[b]

[a]Institute of Informatics of the Slovak Academy of Sciences, Dubravska cesta 9, 845 07 Bratislava, Slovakia
[b]Philosophical Faculty, Comenius University, Gondova 2, 818 01 Bratislava, Slovakia
milan.rusko@savba.sk

The prosody model is one of the most important parts of every speech synthesizer, influencing mainly its naturalness. The intonation contour and durations of phonemes (together with speech quality) bear a great deal of extralinguistic and paralinguistic information contained in the synthesized speech. The features reflecting personality, mood and emotions of the speaker are in strong interaction with those reflecting speech styles. Anyway, the appropriate choice of a prosody model and training material can make it possible to create a special model for every speaking style.

The paper presents our approach to modelling of acoustic parameters of prosody in two different speech styles in Slovak. Our model is based on Classification and regression trees (CARTs). It uses independent CART for phoneme duration and three CARTs for fundamental frequency (F0) at the beginning, centre, and end of each syllable. Two hours of read speech were used for training a model of read speech. The recordings of a puppet player were used to train a model of acted speech. The models were implemented in the Kempelen 2.2 unit selection Slovak speech synthesizer. The listening tests have shown that the models are capable of modeling significant amount of the differences of the two speaking styles.

1 Introduction

In order to produce speech with an acceptable level of naturalness a speech synthesizer should be equipped with a prosody model. The basic acoustic features that have to be predicted are fundamental frequency (F0), durations (quantities) of phonemes, and intensity. In this paper we describe our effort to model the first two of these characteristics in Slovak.

The most frequent approaches to modelling prosody characteristics are either rule based or statistical. The rule based systems utilize the knowledge obtained in former phonetic and phonological studies of the particular spoken language. They refer to the language in general and so they are not speaker dependent.

To preserve individual features of the speaker's voice in speech synthesis it is better to choose some of the data driven methods, from which we selected the CART based model. A previous study showed that the CART model worked well with our read speech database [8] (used for voice *Milan* in speech synthesizer Kempelen). In the present study we wanted to test the potential of the CART model for capturing another speaking style – acted speech.

2 Speech databases

In our experiments we used two speech databases. VoiceDat speech database [5] contains about 2000 sentences pronounced by one speaker and recorded in studio conditions. 1500 sentences belong to so called phonetically rich sentences which were chosen in order to cover as much context dependent phonemes as possible. A Neumann U 87 cardioid condenser microphone with Focusrite Trackmaster pre-amplifier and a hard disk recording system equipped with AARK 20/20+ sound board was used in the sessions. The sampling frequency was 44.1 kHz and resolution was 16 bit. The speech files were later downsampled to 22.05 kHz. The annotation contains orthographic and orthoepic text representation, pointers to phoneme and syllable boundaries as well as pitch marks. This database was used for training of prosody model *Milan* and it is also used during the synthesis process as a source of voice units for the voice *Milan*.

For training a prosody model of acted speaking style we used existing recordings of a puppeteer. We have already analyzed this actor's voice when studying the use of several 'voices' as different characters of the play by a single speaker [7]. A set of eight complete puppet plays played by Bohuslav Anderle (1913-1976) have been recorded containing 21 voices of different characters. Examples of the voices can be found at http://ui.sav.sk/speech/voices.htm. For the present study we selected 165 sentences consisting of speech of two characters – Faust and Don Juan. These two characters were played in a similar way and according to the actor they belong to the same "class of characters". This material was annotated. Phoneme boundaries were labelled using EHMM with consequent manual correction. [9]

Due to low signal quality and small size of the puppeteer speech database, we only used it for creating a prosody model, which is then applied in the unit selection synthesis using the voice *Milan* (database VoiceDat).

3 CART based models

CART - Classification and Regression Trees [1] are a widely used tool for statistical modelling, e.g. [3, 4]. The result of training is function $y = f(x_1, x_2, ..., x_n)$ where y is the output of the model (predictee, in our case duration or F0 value). $(x_1, x_2, ..., x_n)$ is a vector of input features (e.g. type of the sentence, presence of accent in the syllable, position of the syllable, etc.).

The training data consist of vectors $(y, x_1, x_2, ..., x_n)$, where y is a known (measured) value. The result of the training process is the function f, having the form of a decision tree. In our case it is a regressive version of the tree, because the predicted variables are continuous. The classification type of the tree would be used for a discrete output. In the regression tree the final number of classes of y is not known in advance. As we need the tree to give only a limited number of the resulting values, the means of the values in the particular classes are taken as representatives of these classes.

CART consists of nodes with questions like $x_i > k$? (if x_i is a numerical variable) or $x_i \in \{p_1, p_2, ...\}$? (if x_i is a categorical variable). Each question concerns only one input variable and the answer can be either "yes" or "no". Therefore, CARTs are binary trees. The division

continues until some meaningful criteria – determined from the training data – are met. The final nodes (leaves) then represent the resulting value of the predicted variable, valid for a given sequence of answers to the discriminative questions.

The use of the final CART is simple and quick. We have a vector of values of the input variables $(x_1, x_2, ..., x_n)$ and we go through the tree from the root branching according to the evaluation of questions until one of the leaves is reached, which gives then the final value y .

It is possible to make up or modify the CART even manually taking the advantage of its readability and simple conversion to rules (chaining questions leading to some output) and vice versa. In most cases programs for automatic CART building based on statistical evaluation of training data are used. In this work we also followed this approach and we used a publicly available program *wagon* which is a part of Edinburgh Speech Tools [2].

4 Feature description

The choice of features is crucial for modelling duration and intonation with CARTs. Since no text marking is implemented in our Text-to-Speech system yet, we are limited to the variables that could be obtained automatically from the plain text.

4.1 Duration

For the generation of both read and acted models of prosody, phoneme durations in milliseconds were estimated from the two speech databases for every phoneme. For the sake of normalization, the values were expressed in the form of Z-scores. This represents relative duration of the actual phone to the mean duration of all realizations of this phoneme in the database and is evaluated as Z-score = (x - μ)/σ where x is the variable (duration in this case), μ is its mean and σ is the standard deviation of x.

The following features were chosen for the calculation of durations predicting CART:

Phoneme characteristics

- identity of the target phoneme,
- pre-defined articulatory group (or pause) of the preceding and the following phonemes,
- syllabic identity of the phoneme (i.e. its location at the onset, coda, or nucleaus of the syllable);

Syllable characteristics

- the number of phonemes in the syllable,
- the number of phonemes in the coda of the actual syllable and in the onset of the following syllable,
- whether the syllable is accented or not,
- the number of syllables from the last accented one,
- the type of pause before and after the syllable,
- the type of the syllable position in the word (beginning, central, last, single (monosyllabic word));

Word characteristics

- the number of syllables in the word;

Fig. 1 shows an example of generated CART based on duration features.

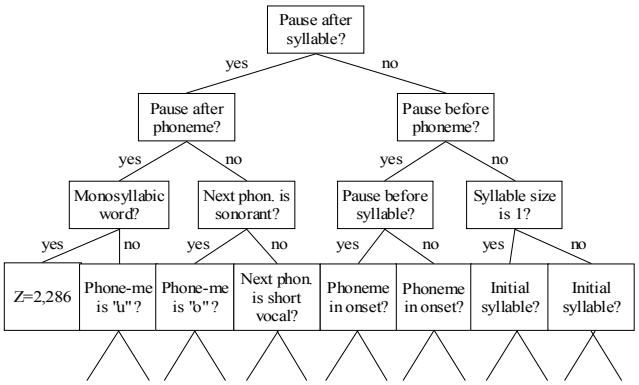


Fig. 1 The first four levels of the CART for phoneme lengths prediction resulting from the training process.

4.2 Intonation

The fundamental frequency is estimated on segments corresponding to the syllables. For each syllable three points are considered: the centre of the syllabic vowel, the beginning and the end of the syllable. As we worked only with male voices where F0 is limited to about 280 Hz, using linear scale to express the values of F0 was sufficient. Nevertheless, the frequencies for model training are taken as relative differences to the mean value of F0 of the respective sentence.

In our CART model the following features were taken into account:

Syllable characteristics

- the number of phonemes in the syllable,
- whether the actual, previous and the following syllables are accented or not,
- the number of syllables from the last accented syllable,
- the type of potential pause before and after the syllable,
- the type of the syllable position in the word (beginning, central, last, single (monosyllabic word)),
- relative position of the syllable within the prosodic word and the phrase;

Word characteristics

- the number of syllables in the word,
- the type of pause before and after the word,
- the type of the actual, previous and following word (content word, conjunction, preposition, etc.),
- the position of the word in the prosodic phrase (first, central, last);

Prosodic word

This is a unit of speech starting with an accented syllable and ending with the last syllable before the following accented one:

- its position in the phrase (beginning, central, final);

Phrase and sentence

- the type of the phrase: either a simple sentence or a part of a compound sentence,
- the type of the sentence (question, declarative, command, unfinished)

5 Implementation

The resulting CART models have been implemented in the Kempelen 2.2 unit selection synthesizer in Slovak using voice *Milan* [6]. Fig. 2 shows the schematic description of the prosody model implementation.

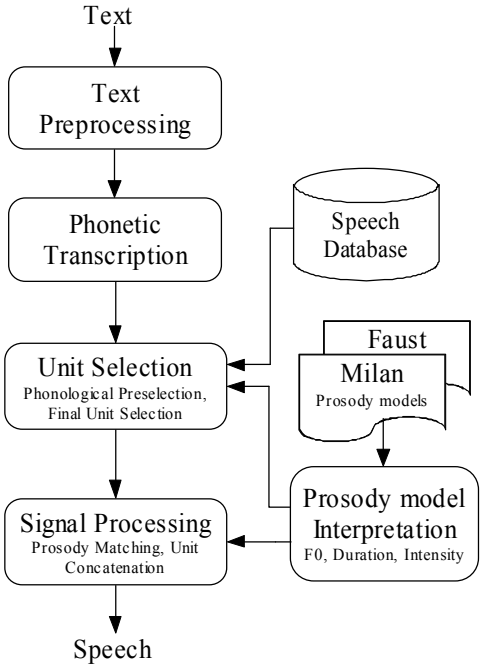


Fig. 2 Schematic diagram of the implementation of the prosody models in the unit selection synthesizer

The prosody model is implemented in the synthesizer in three phases:

Model interpretation

In this phase the values from the CART models together with the information from the sentence (obtained from the text pre-processing) are transformed to a form readable by other parts of the synthesizer (phoneme duration in ms, F0 in Hz);

Unit selection

In this phase the desired values of duration and F0 are taken into account during the selection of the best candidate units from the database. After the concatenation, raw "unit selection" speech signal is obtained.

Signal post-processing

In this phase signal processing (time-warp and pitch-shift) techniques are used to fit the raw speech signal exactly to the model. This phase brings a slight degradation in naturalness, but improves prosody and makes it possible to change overall pitch, pitch range, and speech rate.

6 Results

For preliminary listening experiments we selected 20 utterances synthesized with both prosodic models (read and acted speech) and four listeners. These listening tests were made using our Syn_test program. It enables checking speech concatenated with the prosody model *Milan* (read speech) and *Faust* (acted speech), and the speech after signal post-processing changing the duration and F0 of the selected units exactly to the values desired by the model. It also allows for checking the signal of the particular selected units, their pitch marks and the synthesized signal in a wave editor.

Below we discuss the most salient and consistent observations of the listeners. The first observation is that the phrase-final declination differs in the two models. The intonation (F0 contour) graphs (Fig. 3) show, that in the acted speech model the declination at the end of declarative sentences is not as strong as in the read speech model (see Figs 3b, 3c). Moreover, sometimes the melody even rises in these sentences in the acted speech.

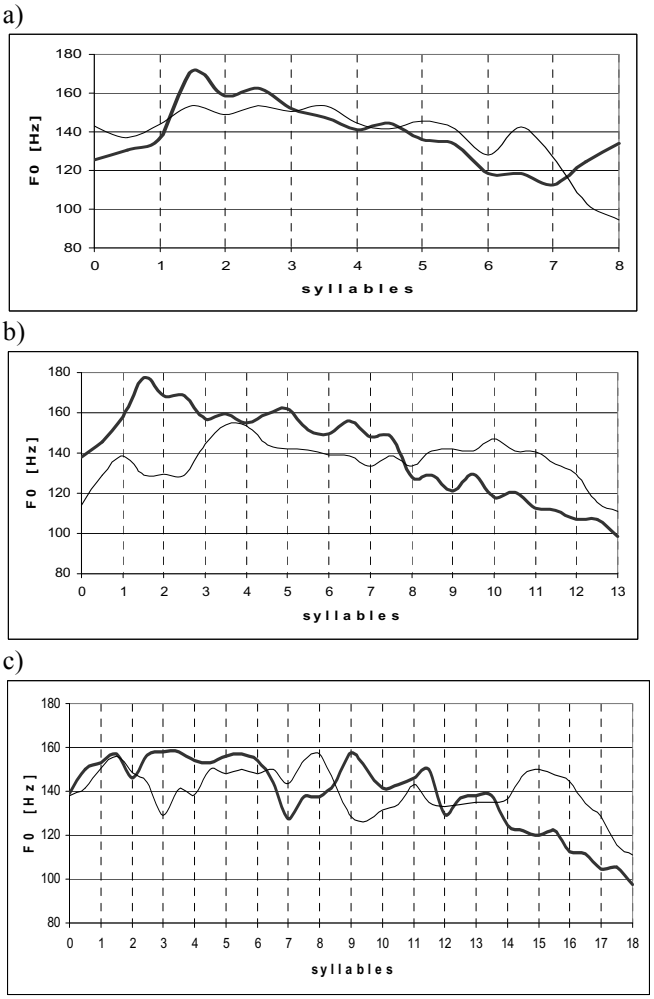


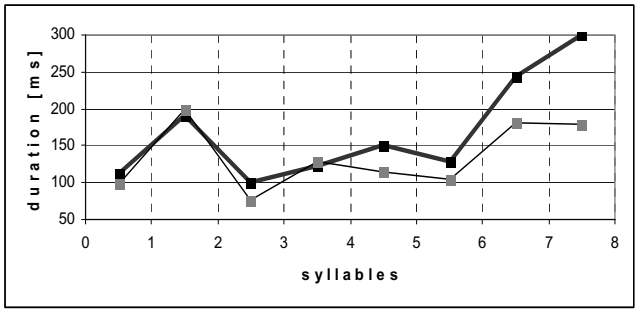
Fig. 3 Comparison of intonation contour of the two models on three sentences: a) "A čo urobíme s ohňom?" (And what shall we do with the fire?); b) "Uvažujeme, že si to zopakujeme." (We consider repeating it.); c) "Neboj sa, nič nebudeš robiť, len budeš dobre rozkazovať." (Don't be afraid, you will have nothing to do but give commands well."). Thick line shows the *Milan* (read) prosody model output and the thin one the *Faust* (acted) model.

A potential explanation for this finding is that in the original actor's sentences the penultimate syllable tends to be made prominent. This feature is different from the reading style where it is always the leftmost syllable of the word that is prominent. Prominence on syllables in Slovak correlates also with increased pitch, which thus may explain the difference in the final declination between the read and acted speech. Furthermore, the puppeteer needs to maintain the attention in the audience; hence, he avoids pragmatic finality signalled by F0 declination that is common for reading style.

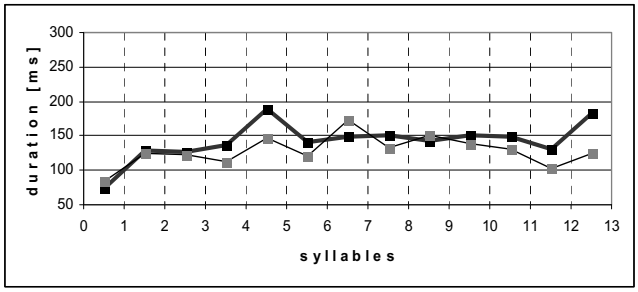
Another observation is the intonation difference in the questions, Fig. 3a. We see that the actor model has in general more flat intonation with sharp F0 declination while the reading speech model produces final rise. Naturally, the training database size (only 165 actor's sentences vs. 1500 sentences of read speech) could be the reason for this difference. Following our discussion about the audience, final declination may be used since questions naturally maintain the attention of the audience until the answer is provided. Furthermore, many questions in the acted speech database are produced as exclamations with falling intonation. Hence, the final declination in the acted speech model in Fig. 3a naturally reflects the training data.

The features that were the most important in the CART prosody model for intonation were the relative position of the syllable within the prosodic word and the phrase, the phrase position in the whole sentence, and the type of the phrase/sentence.

a)



b)



c)

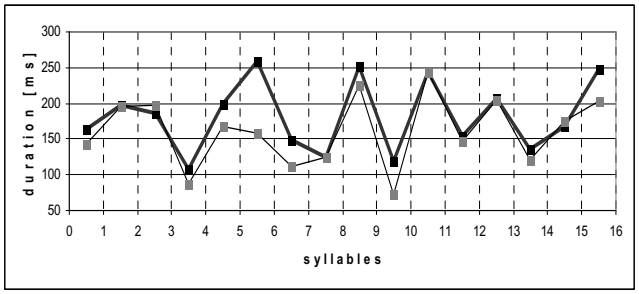


Fig. 4 Syllabic duration comparison of the two models (the same sentences as in Fig. 3).

In terms of duration, we observed that the duration of syllables (calculated from phoneme durations, Fig. 4) in both models is mostly correlated. On average, the acted speech model produces faster speech. However, this is largely due to the final syllable whose duration is much shorter in the acted speech model than in the read speech model. In this sense the presence of final lengthening correlates with final declination: read speech model makes use of both of them more than the acted speech model.

The features that were the most important in the CART prosody model for duration were the position of the syllable in a phrase, the articulatory group of the following phoneme, and the identity of the target phoneme.

After using the two prosodic models for synthesizing individual sentences and their perceptual evaluation, we used the two models for synthesizing longer speech. We used transcriptions of spontaneous dialogues and narratives taken from short stories. We observed that synthetic speech of dialogues created with the acted speech model sounds more natural then with the read speech model. On the other hand, synthetic speech of narratives created with the read speech model was more natural than with acted speech model. Hence, the prosodic features affecting the communicative function of different styles were well preserved from the original training databases.

7 Conclusion

In spite of the fact that standard subjective and objective tests have not been accomplished yet, the preliminary listening tests have shown that the overall speaking style of synthetic speech corresponds to the speaking style of the training database of the selected prosody model. Our results thus suggest that the CART method is suitable for modelling the acoustic characteristics of speaking styles in Slovak. We are currently designing more objective listening experiments to test the validity of preliminary observations.

As expected, it is difficult to preserve personality, mood and/or emotions only with prosody modelling. It is apparent that modelling of the speech quality and other characteristics is necessary.

As this paper is one of our first attempts to model extralinguistic and paralinguistic information, we believe there is a great potential for improvements in our prosodic modelling. We plan to extend our speech database with recordings of other speaking styles and to train several prosody models.

Acknowledgments

This work was supported in part by the of the Ministry of Education of the Slovak Republic, Scientific Grant Agency project number 2/0138/08 Applied Research project number AV 4/0006/07 and by the European Education, Audiovisual and Culture Executive Agency LLP project EURONOUNCE.

This project has been funded with support from the European Commission. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

References

[1] Breiman, Friedman, Stone, Olshen, "Classification and Regression Trees", Chapman Hall, New York, USA, 1984.

[2] Speech tools manual, http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0

[3] Batůšek, R., "A Duration Model for Czech Text-to-Speech Synthesis", Proceedings of Speech Prosody 2002, Aix-en-Provence, France.

[4] Cosi P., Avesani C., Tesser F., Gretter R., Pianesi F., "On the Use of Cart-Tree for Prosodic Predictions in the Italian Festival TTS", In: Cosi P., Magno Caldognetto E., Zamboni A. (Ed.), in Voce, Canto, Parlato – Studi in onore di Franco Ferrero (pp. 73-81). UNIPRESS, Padova, 2003.

[5] Rusko M., Darjaa S., Trnka M., Cerňak M., "Slovak Speech Database for Experiments and Application Building in Unit-Selection Speech Synthesis", In: Proceedings of Text, Speech and Dialogue, TSD 2004, Brno, Czech Republic, pp. 457 – 464.

[6] Rusko M., Trnka M., Darjaa S., "Slovak TTS - From Rule Based to Unit Selection", Proceedings of the international conference Language Technologies IS-LTC 2006, Ljubljana, Slovenia. ISSN 1581-9973, pp. 261-266.

[7] Rusko M., Hamar J., "Character Identity Expression in Vocal Performance of Traditional Puppeteers", In: Text, Speech and Dialogue, 9th International Conference, TSD 2006, Brno, Czech Republic. LNAI 4188. ISBN 978-3-540-39090-1, pp. 509-516.

[8] Rusko M., Trnka M., Darjaa S., Kováč R. "Modelling acoustic parameters of prosody in Slovak using Classification and Regression Trees", In: Human Language Technologies as a Challenge for Computer Science and Linguistics - Proceedings. Poznań, Poland, 2007. ISBN 978-83-7177-407-2, pp. 231-235.

[9] Black A.W., Lenzo K. A. "Building Synthetic Voices " http://festvox.org/bsv/