Naturalness of Prosody Models[1]

Jan Romportl, Daniel Tihelka, Jindřich Matoušek

E-mail: rompi@kky.zcu.cz, dtihelka@kky.zcu.cz, jmatouse@kky.zcu.cz

1. INTRODUCTION

The text-to-speech system ARTIC (developed at the Department of Cybernetics, University of West Bohemia in Pilsen) basically includes two different prosody models: Rule-Based (RB) and Data-Driven (DD). Unlike many other speech-technology problems, the final evaluation of synthetic speech *naturalness* must be carried out in a "subjective" (or "inter-subjective") way – this because of its very nature. Our article presents results of two tests evaluating naturalness of prosody generated by the aforementioned RB and DD models, together with a very brief description of both models.

2. RULE-BASED MODEL

The RB model is based on applications of linguistically motivated rules. It places simplified models of cadences (intonation patterns of prosodic words; by "prosodic word" we mean an intonation foot) and prosodemes (abstract linguistic functionally relevant units, similar to phonemes or morphemes, but in the level of suprasegmental acoustic speech characteristics). This model is intended to be able to carry out suprasegmental modulations of speech melody (e.g. fundamental frequency – F0 contour), intensity (e.g. volume) and timing (e.g. phone duration).

Quantitative characteristic of the rules is represented by a set of 21 experimentally set up parameters, such as F0 value of the stressed vowel of a prosodic word relatively to the initial F0 value of the prosodic word, initial F0 value, stressed syllable volume and duration coefficients, etc. Structurally the model consists of the inventory of 14 cadences and 4 prosodemes (which is very simplifying). For a typical F0 contour produced by this model see Figure 1. Detailed description of it can be found in [1].

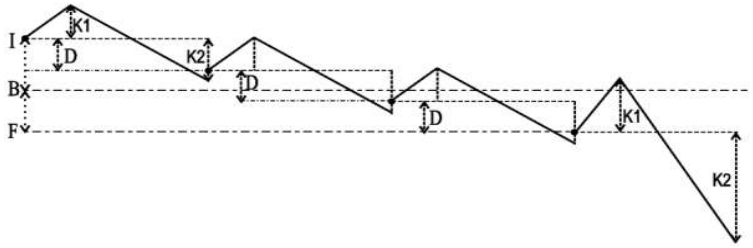

Figure 1: Typical F0 produced by the RB model

1 Research supported by Ministry of Education MSM235200004

3. DATA-DRIVEN MODEL

This approach is being developed due to the limitations (concerning the speech naturalness) of the rule-based approach. The idea is to set up model parameters automatically, using real speech data from a corpus.

The currently used corpus consists of 5,000 sentences by a female speaker. The speech data are segmented into prosodic words by the same method used in the rule-based approach. This way we have about 50,000 prosodic words with detailed representation of F0 shapes of these words. The cadence inventory is then created using an agglomerative clustering algorithm which creates *n* clusters whose centroids (or other representants) are considered to be cadences representing the variability of intonation patterns over prosodic words. We experiment with *n* ranging from *10* up to *200*. The current application of the model is aimed only at the F0 contour; timing and volume modifications are still produced by the rules. The extensive description of this model is in [1] and [2].

4. PROSODY NATURALNESS EVALUATION

We have undertaken experiments aimed at deciding which approach produces better prosody for synthetic speech in terms of its naturalness. This article describes two of these tests: direct comparison of RB and DD models (we will refer to it as the Test 1), and indirect naturalness evaluation using MOS tests (Test 2).

4.1. Test 1

The test was realized as a listening test with 16 respondents, most of which were university students and staff. The test consisted of 10 different sentences with various length, each sentence was synthesized using RB and DD model and respondents were asked to designate its variant, which – according to their opinion – sounds more like a "human-spoken", concerning its prosody qualities. The test was carried out twice – separately for a female and a male voice. Figures 2 and 3 show the absolute preferences in favour of RB and DD for the female and male voice respectively.

According to the results the test has shown following conclusions:

- DD model is clearly preferred

- laic listeners tend to prefer rather "machine-like" sounding RB version when otherwise more "human-like" DD variant exhibits even a single intonation anomality (perceived "uncomfortably" by the listeners)

- TTS-concerned listeners on the other hand easily tolerate such problems in favour of the DD model

- the DD score is even better for the male voice, although the DD model has been set up using female speech data (!) - it supports the assumption male voices are more suitable for synthetic prosody than female ones, perhaps due to their lower absolute values of the fundamental frequency (e.g. higher distance between F0 and other formants which results in less unwanted artefacts in the phone quality caused by modifications to the demanded F0)

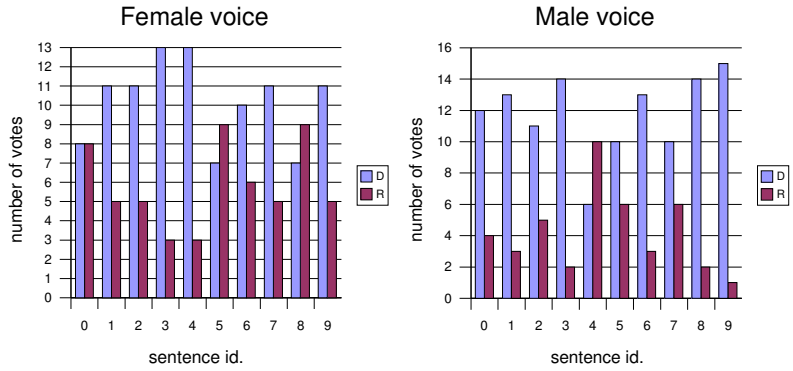

Figure 2: Sentence preferences

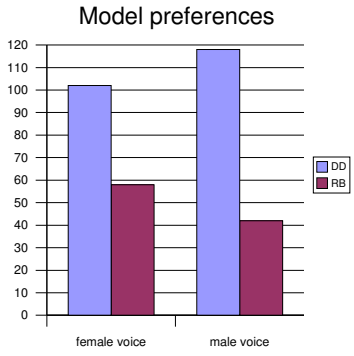

Figure 3: Overall DD and RB preferences

4.2. Test 2

This test indirectly compares RB and DD models by measuring the inter-subjective criteria of the prosody quality using MOS tests (mean opinion score). We used the same participants as in the Test 1, yet with different sentences: 12 different sentences were synthesized as follows – 3 monotonous (further denoted as *m*), 3 rule-based (*r*), 3 data-driven (*d*), 3 real "implanted" (*i*, synthesized with the F0 taken from the female

speech corpus and with the same sentence as occurred in the corpus). Each listener received these sentences in a random order (obviously listeners were not told which method was used to generate a particular sentence) and was asked to give each sentence a mark from the scale 1 – 5 (1 stands for best, 5 for worst) according to his/her subjective opinion. No prior "calibration" (i.e. examples of good or bad sentences) was presented to the listeners since we wanted them to express their own understanding of what is (un)naturally sounding prosody. This test was carried out only for the female voice.

Moreover, the results were evaluated in two different variants: simple (S-MOS) – the values of the marks were taken as they were given; normalised (N-MOS) – the marks from each listener were normalised (using linear transformation) so as to cover the whole scale 1 – 5 (for example when a listener gave marks only from the scale 2 – 4, they were linearly transformed to the scale 1 – 5). S-MOS are thus more suitable for judging the overall view of how natural the generated prosody is (in terms of listeners' opinion), while N-MOS is better for comparison between the prosody models. Figures 4 and 5 show the results.

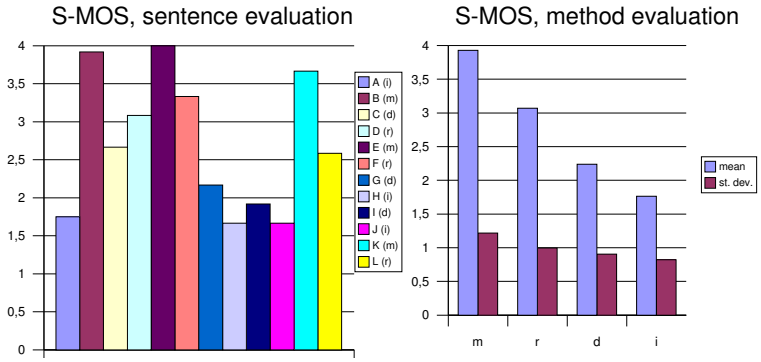

Figure 4: Results of the S-MOS test

The Test 2 offers these conclusions:

- considerably higher naturalness of DD confirmed

- DD is not evaluated much worse than than the real intonation

- real intonation is evaluated almost by the mark 2; it might point out that listeners cannot fully separate the segmental from the suprasegmental qualities of the synthetic speech, even though they are instructed to do so

- monotonous is often evaluated as quite natural sounding

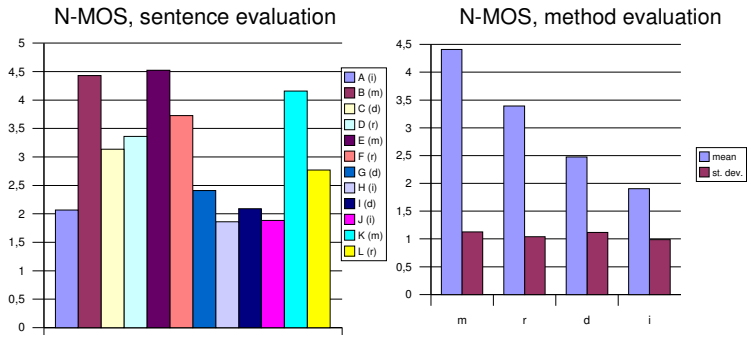

Figure 5: Results of the N-MOS test

5. OVERAL CONCLUSIONS

The results of the tests presented in this article show that even the very first version of our Data-Driven model is very encouraging concerning its performance in a text-to-speech system. Another interesting result is the fact that a model of the female intonation can be (under some extent and conditions) used also for a male voice; more generally – such a DD model can be set up with speech data of one speaker and used with a voice of another one. And last but not least, we have found out that more relaxed (e.g. rather monotonous) synthetic speech intonation is often perceived far better than the excessive one.

REFERENCES

[1] Romportl, J.: Generování prozodie z textu pro účely syntézy řeči (Generating Prosody from Text for Speech Synthesis Purposes). Západočeská univerzita v Plzni (University of West Bohemia in Pilsen), Pilsen (2004).

[2] Romportl, J., Matoušek, J., Tihelka, D.: Advanced Prosody Modelling. In Proceedings of 7th International Conference on Text, Speech and Dialogue, TSD 2004. Berlin : Springer-Verlag, 2004. p. 441-447.