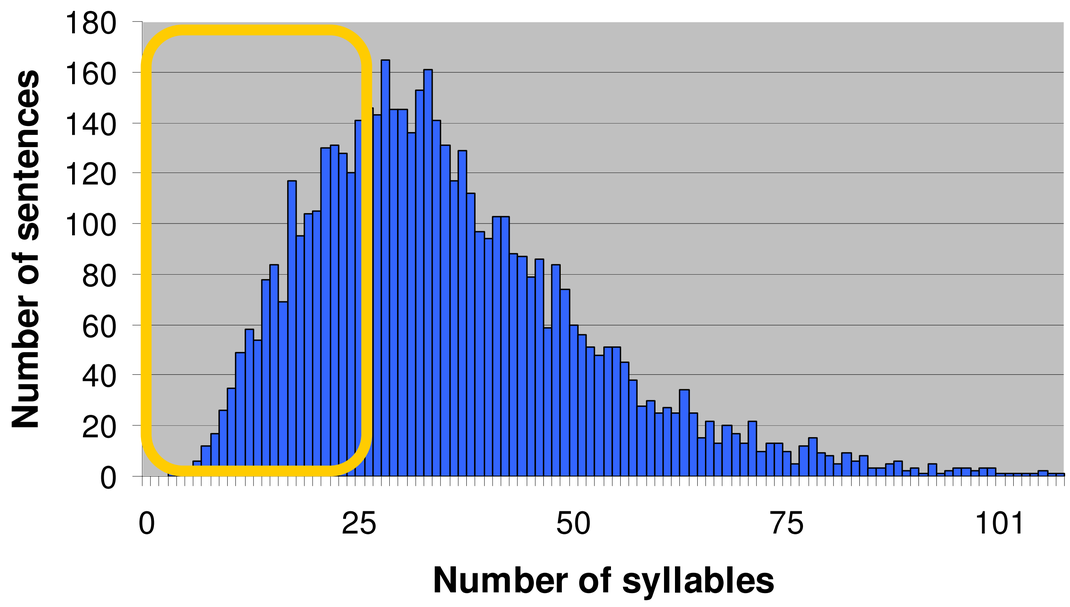
INCREASING PROSODIC VARIABILITY OF TEXT-TO-SPEECH SYNTHESIZERS

1. Introduction

- Current speech synthesis systems are still recognized as non-human when synthesizing extended passages.
- The prosody component of TTS systems is designed to generate the intonation of formal text, and lacks the variability of natural speech.
- The goal of this work is to generate more natural prosody and to introduce variability over successive sentences.

- Difference from other approaches (Dong, Lua (2000); Raux, Black (2003); van Santen et al. (2005)):
  - no full prosodic model, emphasis on variability,
  - single layer approach to a whole prosodic phrase, the minimalist selection procedure using word and syllable counts and a priori stress information.
- Method might directly work for languages with fixed stress position, like e.g. Hungarian, Finnish and can be extended to other languages by a proper sentence (prosodic phrase) similarity measure.



Histogram of the number of syllables in the database of natural weather forecast sentences (based on 5200 sentences). The maximum length of sentence selection was limited to 25 syllables in order to better approximate standard sentence lengths of Hungarian.

2. Methodology

- Purpose: to improve prosody generation of a TTS in repetitive similar (or same) utterances (e.g. Good morning.).

- For a given declarative input sentence first a standard rule-based prosody is applied.

- A database of natural sample sentences is searched for sentences having similar syllabic structure to the input (simplified definition of the syllabic structure of a sentence: the number of words in the sentence and the number of syllables in each word).

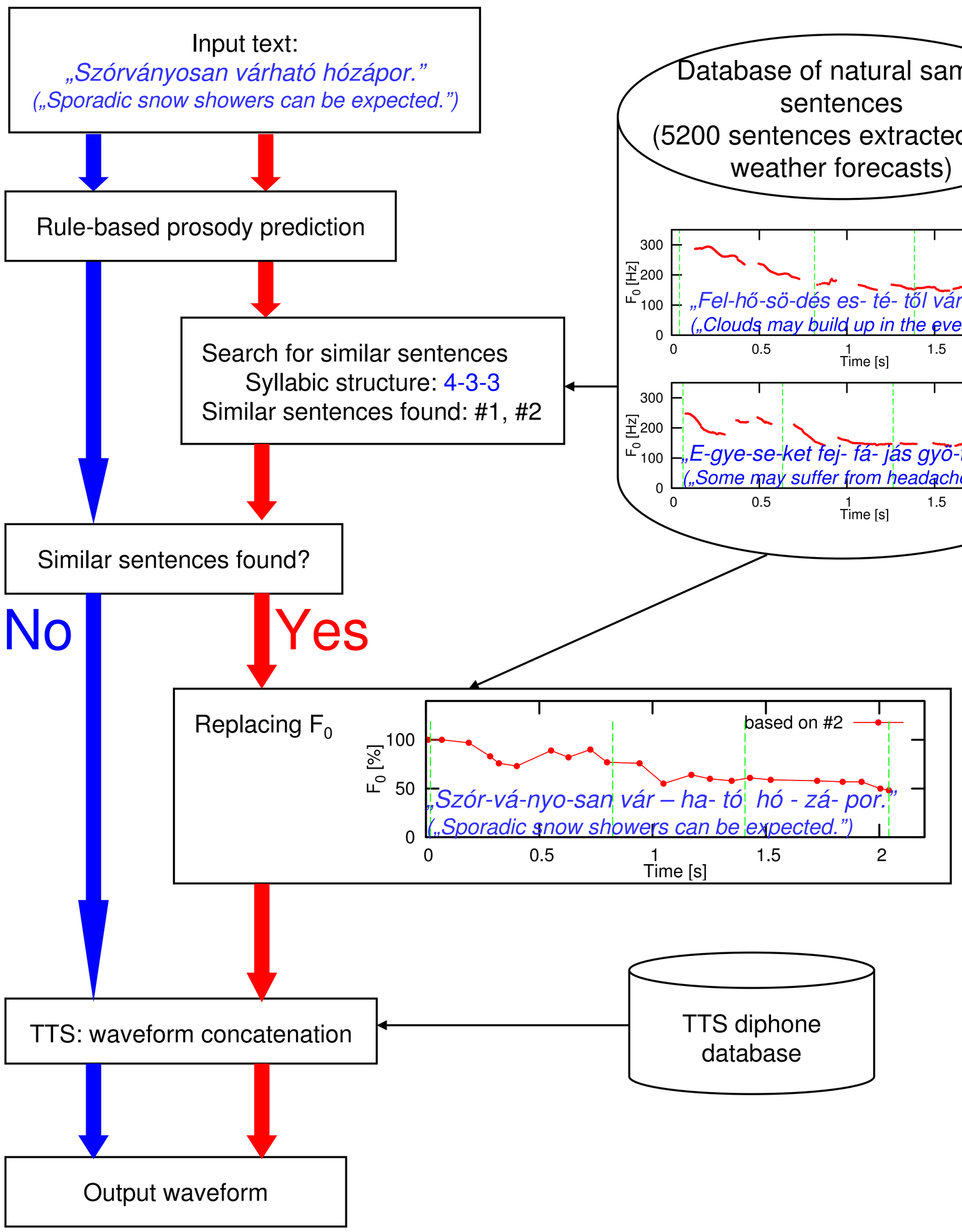
- If no similar sentence is found, the standard rule-based F0 contour is used.

- If similar sentences are found, one of them is selected randomly.

- The prosody of the randomly selected natural sentence (in this experiment only F_0) is used as a target to generate the prosody of the synthetic one by syllable-based time warping.

- The output waveform is produced by a diphone TTS synthesizer.

- The method allows us to reduce the monotony of utterances, even in case of the repetition of the same sentence.



Géza Németh, Márk Fék, Tamás Gábor Csapó
Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Hungary

{nemeth, fek, csapo}@tmit.bme.hu

3. Experiments

- Goals:
  - to compare the F_0 copying method to the standard rule-based solution,
  - to evaluate different natural-based F_0 variants.

- 6 sentence groups with matching syllabic structure were selected including some variants for each sentence in the group:
  - one with a rule-based F_0 contour,
  - 1-3 with an F_0 contour copied from sentences in the group.

4. Tests and results

- 13 sentences were selected and sentence pairs were created of them for a 3 level comparison (#1 more natural than #2, equal or #2 more natural than #1).
- The results of a web-based test from groups of at least 20 listeners (altogether 208 listeners) were evaluated.

5. Conclusions

- This initial study was successful based on the perceptual tests: listeners preferred the F_0 copied variants over versions with rule-based pitch contour.
- The new method requires better signal processing solutions.

6. Further plans

- Extend this work with the timing and intensity features of prosody,
- Improve signal processing,
- Extend domain coverage.
- By refining the similarity measure results may be further improved.
- Applying to other languages:
  - fixed stress: method works without modification: e.g. Finnish, Polish
  - varying stress: by a proper sentence (prosodic phrase) similarity measure: e.g. English.

7. Acknowledgements

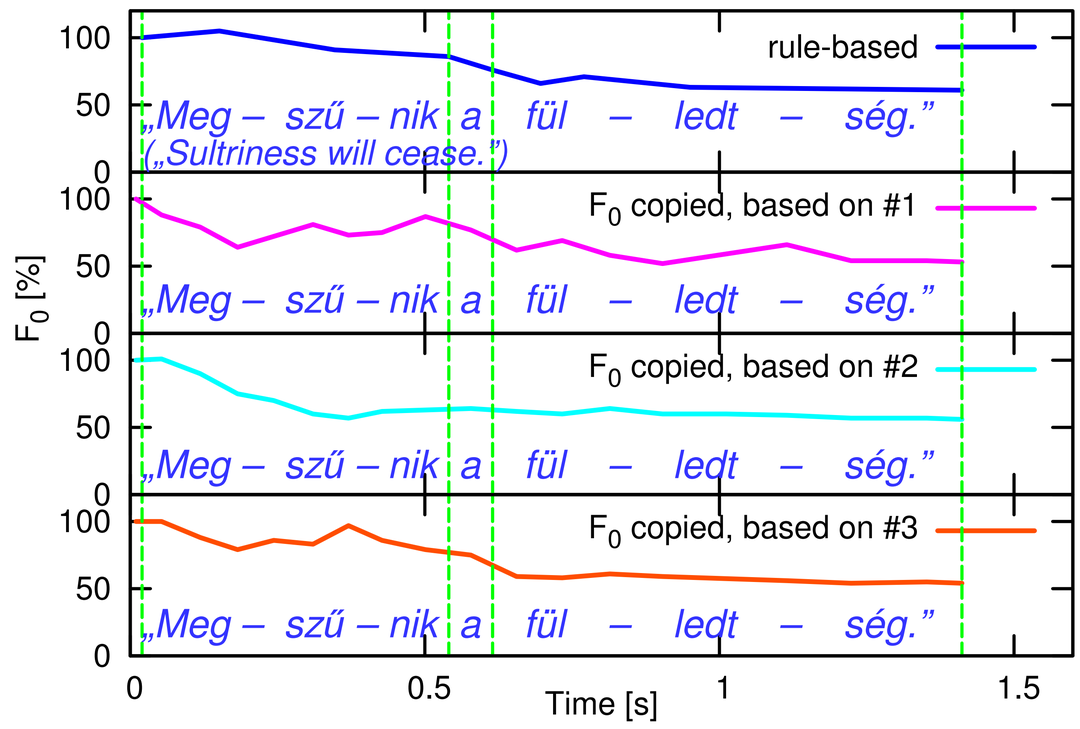
- We thank all listeners for participating in the subjective tests.
- The research presented in the paper was partly supported by the Hungarian National Office for Research and Technology (NKFP 2/034/2004 and NAP 00736/2005).

8. Key references

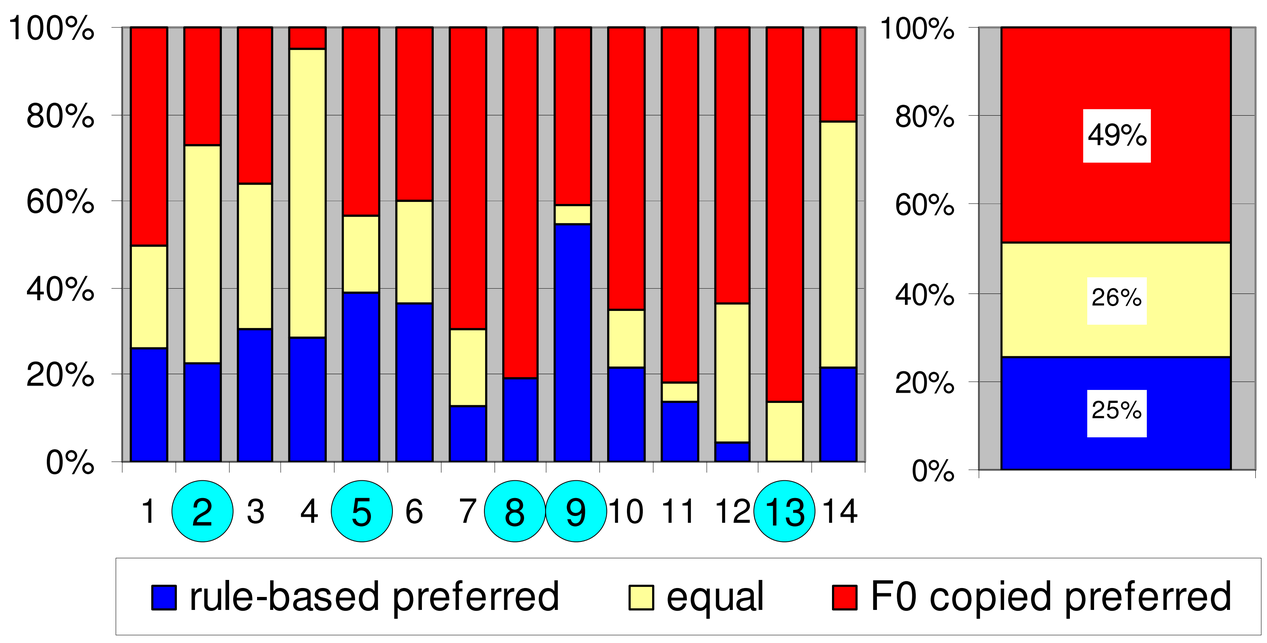
Dong, M., Lua, K. T. (2000) “An Example-based Approach for Prosody Generation in Chinese Speech Synthesis”, ISCSLP, Beijing, pp. 303-307.

Raux, A., Black, A. (2003) “A Unit Selection Approach to F0 Modeling and its Application to Emphasis”, ASRU, pp.700-705.

Van Santen, J., Kain, A., Klabbers, E., and Mishra, T. (2005) "Synthesis of Prosody using Multi-level Unit Sequences”, Speech Communication, Volume 46, Issues 3-4, pp. 365-375.

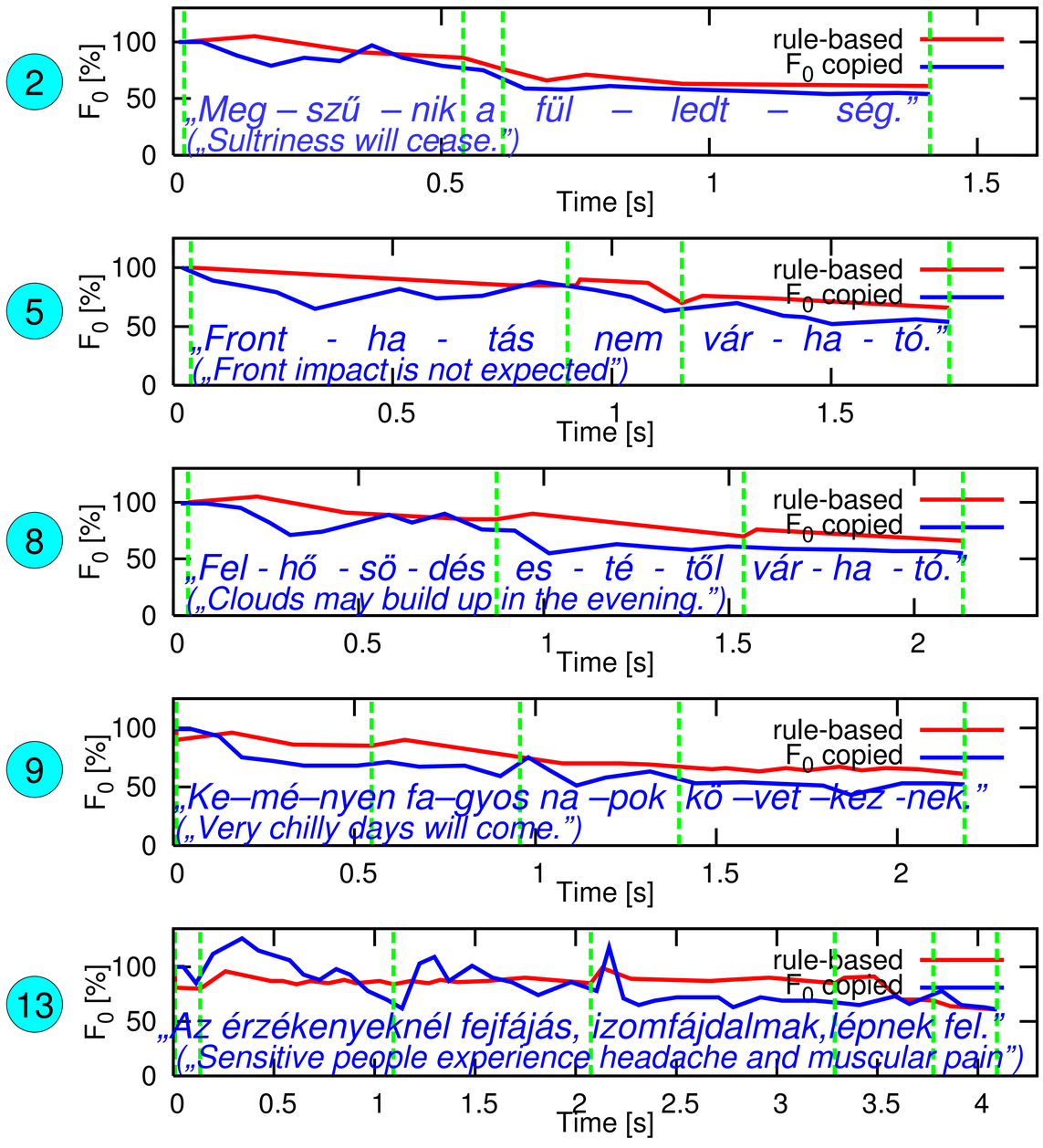


Pitch contour of a sentence in four styles



Results of the comparisons between rule-based and semantically different F_0 copied variants of sentences

Comparisons summarized



Pitch contour of 5 sentences in two styles