RULE BASED SOUNDS DURATION MODEL FOR THE CZECH TTS SYSTEM

Petr Horák

Institute of Radio Engineering and Electronics, Academy of Sciences of the Czech Republic
horak@ure.cas.cz

Abstract: A phoneme duration model is a standard part of current text-to-speech (TTS) synthesizers. Our contemporary TTS systems have been using the relative duration modeling of separate speech units (diphones or triphones). The information about sounds borders is not available during speech synthesis. Only the information about borders of speech units is available. This paper deals with extension of the relative duration based TTS system. The aim is the TTS system with absolute duration modeling of separate sounds and the possibility of the using MROLA compatible Czech voices in the Epos TTS system. This task consists of extension of speech inventory about sounds margins information and extension of TTS system about separate sounds duration modeling. The main motivation for this work lay in the poor timing of our Czech triphone synthesis with relative duration modeling.

1 Introduction

The current stable version of TTS system Epos 2.4 has been using our proprietary interface between text processing and speech synthesis parts. The communication unit of this interface is segment (basic speech unit). Every segment is a quadruple of segment number, assigned frequency (pitch), intensity (volume) and time factor (speed) encoded as 32-bit little endian integer. Each segment has only three prosody parameters [1]. You can see this interface on Fig. 1.

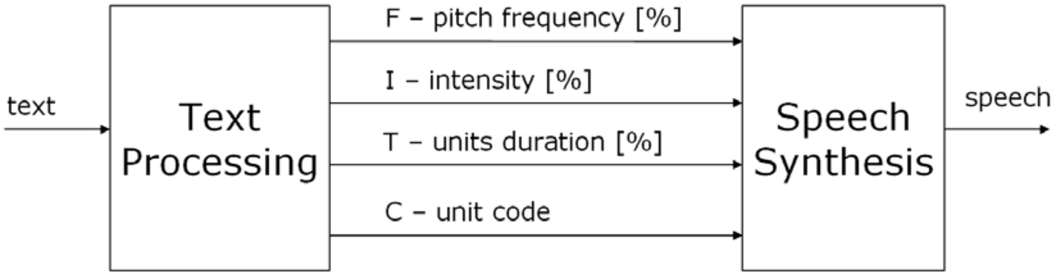


Figure 1 - Internal structure of Epos 2.4.

The Epos 2.4 is from these reasons inapplicable for prosody research and, for example making one syllable questions. Next disadvantage is wrong prosody modeling with triphone synthesis. Usual triphones are two times longer than usual diphones. One trinity of prosody parameters to one triphone faces to unnatural low quality prosody [2][3]. On Fig. 2 we can see waveform of Czech word “potok” (stream) with diphones and triphones. There are visible that this word consists of the six diphones but only three triphones.

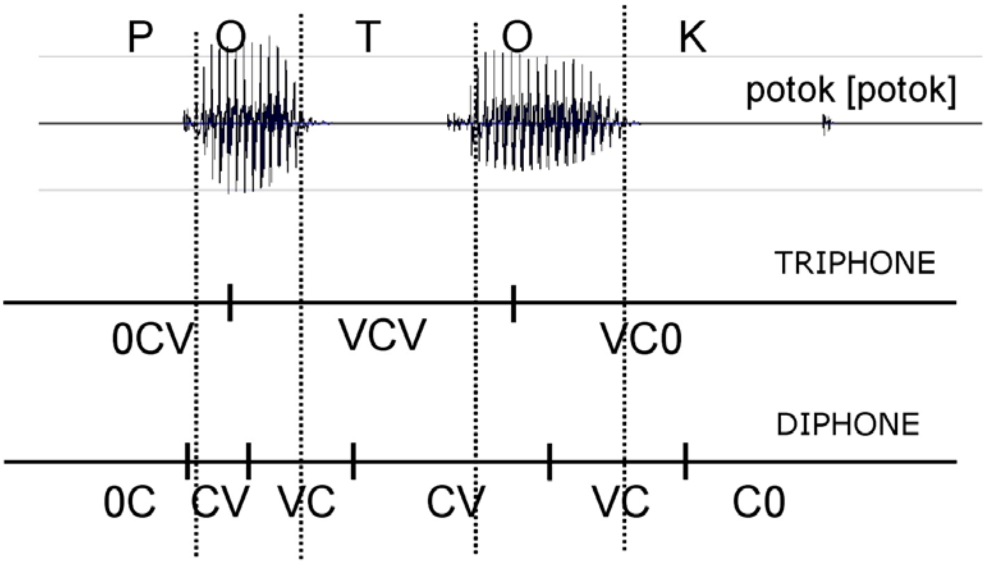


Figure 2 - Structure of Czech word “potok” (stream)..

The development of the new version of TTS system Epos had been started in 2000. This new version use the standard communication interface on the sounds level with multiple prosody points possibility. The new text oriented communication protocol is based on the MBROLA standard with extension of intensity modeling. This interface you can see on Fig. 3.

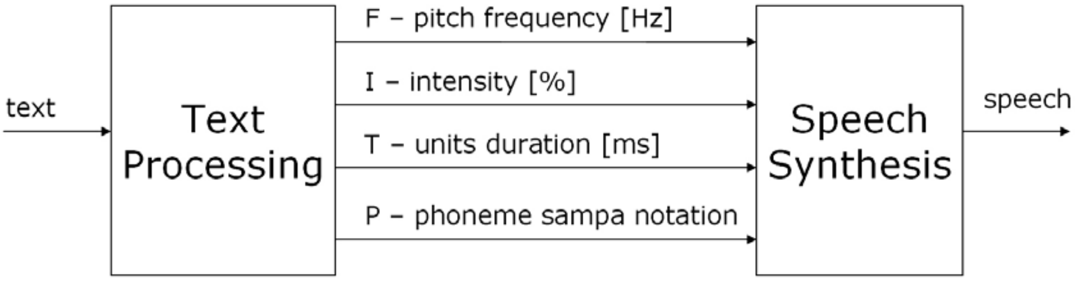
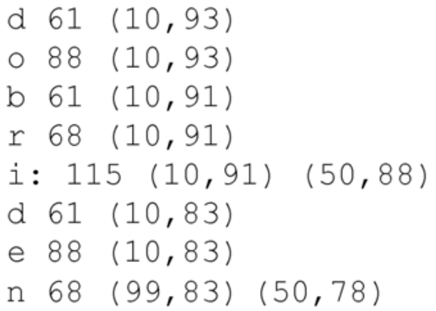


Figure 3 - Internal structure of Epos 2.5.

The prime MBROLA protocol which had been introduced by the MBROLA synthesizer development team allows modeling of sounds duration and modeling of pitch frequency via multiple prosody points for each sound. Our extension of this protocol called Speech Synthesizer Input Format (SSIF) enables intensity modeling in every prosody point. This extension was discussed with MBROLA development team [5].

The Speech Synthesizer Input Format is line oriented, each line corresponding to a single phone. The line contains several white space separated components. The first component is the SAMPA notation of the phone and the second component is its duration in milliseconds.

Subsequent components are prosody points. Each prosody point is enclosed in parentheses and consists of two or three integers separated by commas. The first value locates the prosody point within the phone per cent (e.g. the value of 99 corresponds to just before the end of the phone), the second value indicates the desired pitch at that prosody point (the value of 100 indicates the default pitch) and the third value, which is not currently supported by MBROLA, and which is optional, indicates the intensity at that point. The example of communication via SSIF protocol is illustrated on Fig. 4.



```
d 61 (10,93)
o 88 (10,93)
b 61 (10,91)
r 68 (10,91)
i: 115 (10,91) (50,88)
d 61 (10,83)
e 88 (10,83)
n 68 (99,83) (50,78)
```

Figure 4 - Epos 2.5 communication protocol SSIF.

Using of the MBROLA compatible protocol allows connecting MBROLA speech synthesizers with Epos 2.5.

The male speech inventory “machač” was extended of the sounds borders information using manual labeling with SpeechStudio software system. The SpeechStudio system was extended of sounds duration statistics processing. First sounds duration rules based on timing statistics have been created. The automatic labeling of sounds borders via the DTW algorithm [4] with manual corrections in the SpeechStudio software system was applied too.

2 Conclusions

Up to now we have created basic tool for sounds duration research. Future goals are creation of advanced duration modeling rules and duration modeling using of artificial neural networks.

Acknowledgements

This research has been supported by the Grant Agency of the Czech Republic (GA102/02/0124 “Voice Technologies for Support of Information Society”) and by the Ministry of Education, Youth and Sports of the Czech Republic (OC 277.001 “Transformation of Segmental and Suprasegmental Speech Models”).

References

[1] Hanika, J., Horák, P., “Epos - A New Approach to the Speech Synthesis” In: Proceedings of the First Workshop on Text, Speech and Dialogue - TSD’98, Brno, Czech Republic, September 23–26, 1998, pp. 51–54.

[2] Hanika, J., Horák, P., Text to Speech Control Protocol In: Proc. of the Int. Conf. Eurospeech’99, Budapest, Hungary, September 5–9, 1999, Vol. 5, pp. 2143–2146.

[3] Hanika, J., Horák, P., Depedences and Independences of Text-to-Speech, In: Hans-Walter Wodarz, editor, Forum Phoneticum 69, Frankfurt Am Main, 2000, pp. 27–40.

[4] Horák, P.: Automatic Speech Segmentation Based on DTW with the Application of the Czech TTS System. In: Improvements in Speech Synthesis, Ed. by E. Keller, G. Bailly, A. Monaghan, J. Terken & M. Huckwale, John Wiley & Sons, Ltd., 2002, pp. 328–338.

[5] Hanika, J., Epos on-line documentation. http://epos.ure.cas.cz/