USAGE OF TD PSOLA ALGORITHM IN SLOVAK SPEECH SYNTHESIS BASED ON THE EMU DATABASE SYSTEM

Andrej VRÁBEL, Bachelor Degree Programme (4)
Dept. of Telecommunications, FEI STU Bratislava
E-mail: av2@post.sk

Supervised by: Dr. Gregor Rozinaj

ABSTRACT

Speech synthesis is the process of generating acoustical signal, represented by a sequence of vowels, with particular factual and grammatical meaning. One of the possibilities for speech synthesis is concatenate synthesis, in which prepared partitions of human speech are concatenated, by using convenient algorithms.

The use of the EMU database system, in phoneme based speech synthesis has been reviewed. This paper continues research of this area. At the beginning, it analyses the TD PSOLA algorithm. Then, it solves the problem of adaptation the EMU system for diphone based text-to-speech synthesis. Finally, it introduces the possibility of implementation the algorithm in the EMU system.

1 INTRODUCTION

The TD PSOLA (Time-Domain Pitch Synchronous OverLap Add) algorithm enables modification of speech signal in time domain, by using prepared diphone database. Because of its high computational efficiency, it brings a great advance in text-to-speech synthesis.

2 PITCH AND DURATION MODIFICATION OF SYNTHESIZED UNITS ACCORDING TO THE TD PSOLA ALGORITHM

The synthesis based on TD PSOLA algorithm come out of human speech signal, divided into segments, called diphones. We are able to achieve synthesized speech by concatenating these segments. Moreover, the algorithm enables to modify the pitch and duration of the speech. For periodic signals, we are able to modify the pitch by changing distance among the periods, and duration by adding or omitting some of them. For non-periodic signals, we are only able to modify duration of particular parts of the signal.

Provided infinite periodic signal, we are able to shift period from original T_0 to required T, by summing windowed data $s_i(n)$, originated from x(n) signal.

$$s_i(n) = x(n).w(n - i.T_0) \qquad \text{s(n)} = \sum_{i=-\infty}^{\infty} s_i(n - i(T - T_0)) \qquad (1)$$

The samples $s_i(n)$ only differ from zero on an interval dependent on recovering factor F_R, defined as a ratio of size L of the analysis window w(n) by the pitch period.

$$F_R = L / T_0 \qquad (2)$$

In practice, we choose $F_R \approx 2$, when the spectrum of $s_i(n)$ signal approximates the spectrum of s(n) signal. Then the concatenation process changes the pitch without affecting the formants' frequencies. The use of different recovering factor causes strong degradation of synthesized speech, e.g. buzzing or effect of metal voice.

3 THE CONCEPT OF SYNTHESIZER BASED ON THE EMU DATABASE

The conceptual scheme of the diphone synthesizer is shown on the figure 1. There are three basic phases of synthesis. In the first phase, it is necessary to create a database to store samples of human language. Then, the particular segments of corpus, such as phonemes, diphones, periods, etc. are automatically recognized. This step is done by the special recognizing software, the part of our system. Time marks for the particular segments of corpus, the output of the software, are stored in the database.

The second phase is the speech synthesis. User writes text, he wants to synthesize with desired frequency. Firstly, the text is analysed. Then the database is searched and according to the rules of the language, appropriate diphones are marked. Finally, they are synthesized according to the TD PSOLA algorithm. Synthesized speech is stored in the database for further processing (the third phase), or it can be played over an output device.

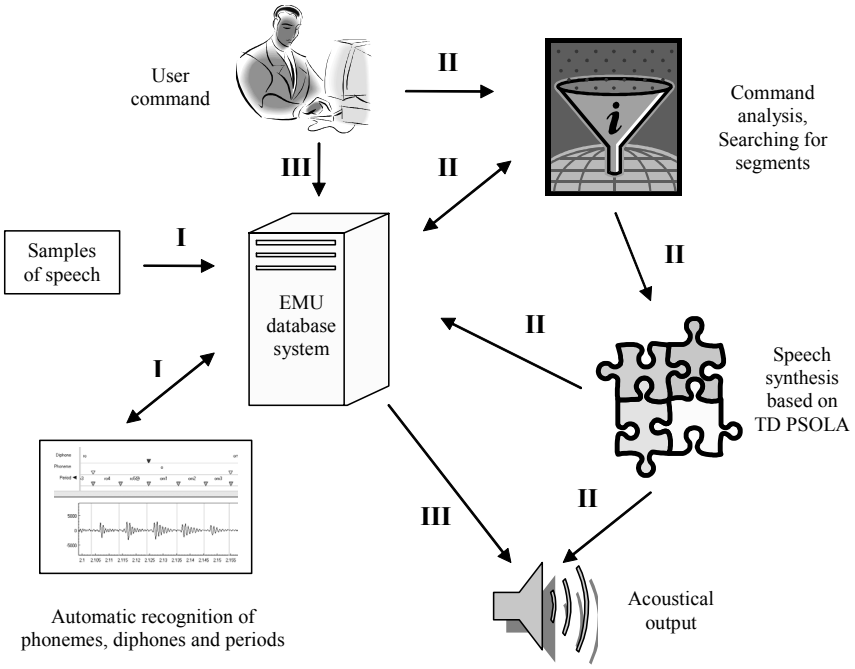

Fig. 1: *Conceptual scheme of diphone synthesizer*

4 THE EMU DIPHONE DATABASE

4.1 DATABASE TEMPLATE

The first step in database creation process is the creation of database template. Ours one was created in the *EMU Template Editor*. There were three basic hierarchical levels chosen: phoneme, diphone and period level.

4.2 HIERARCHICAL LEVELS ANNOTATION

Hierarchical levels annotation is a process of association hierarchical time marks with acoustical signal. After loading speech samples into database had been completed, annotation of phoneme and diphone level could be done. In the period level, periodic parts of the signal were divided into periods. The problem of dividing non-periodic parts was solved quite simply. They were divided to the sections, during as long as the periodic sections do. Sections in hierarchical levels are marked in SAMPA computer alphabet.

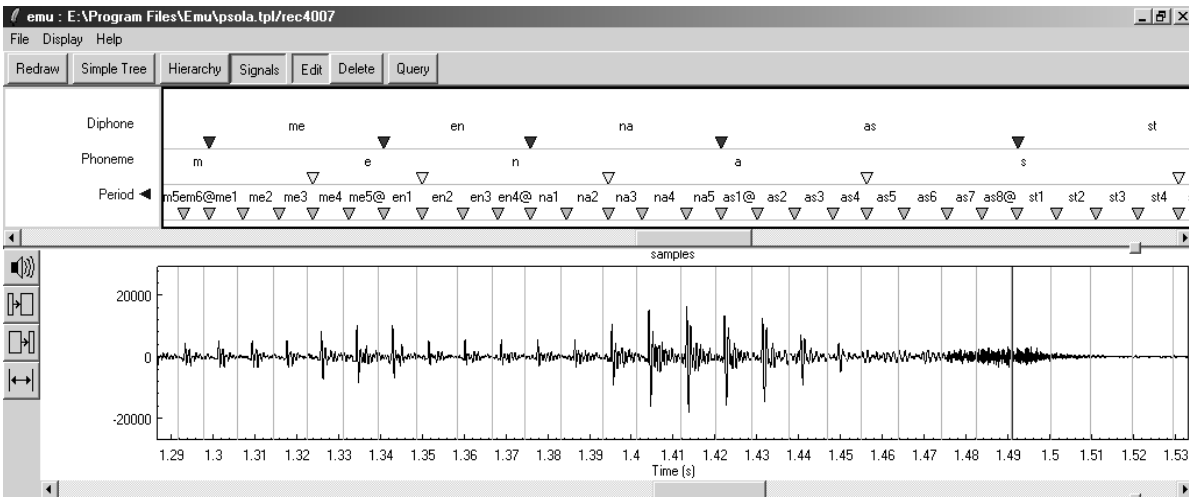

Fig. 2: *Annotation of hierarchical levels in speech signal*

5 IMPLEMENTATION

Both partial algorithms of TD PSOLA are implemented in TCL language, which enables programmer to manipulate with the EMU core libraries.

5.1 DURATION MODIFICATION OF SYNTHESIZED SIGNAL

According to input parameters (diphones divided into periods or non-periodical segments, number of segments to add or omit) a new list of segments is generated. If a user wants to increase the duration of synthesised text, particular periods (marked '@') are being duplicated in the new list. On the other hand, if he wants to shorten the duration of synthesised text, particular periods in the new list are omitted.

5.2 PITCH MODIFICATION OF SYNTHESIZED SIGNAL

The algorithms' input parameters are time marks for synthesised diphones, type of analysis window, and coefficient for relative change of pitch. For each period of diphone:

- the maximum value of amplitude is found, the corresponding time is marked **t**
- the segment in time interval <**t** - $\mathbf{T_0}$; **t** + $\mathbf{T_0}$> is taken
- it is windowed by a window, length 2.$\mathbf{T_0}$**,** Rectangle, Bartlett's, Hamming's, Hanning's and Blackman's window is implemented
- non-overlapping part of segment is saved into the output file directly
- overlapping part of segment enters new cycle, where it is summed with corresponding part of next segment and saved into output file

Because the length of extracted segments is 2.$\mathbf{T_0}$, it is necessary to use 50 % overlap to achieve the original duration of speech. The use of a greater overlap provides the increase of pitch. On the other hand, smaller overlap causes decrease of pitch.

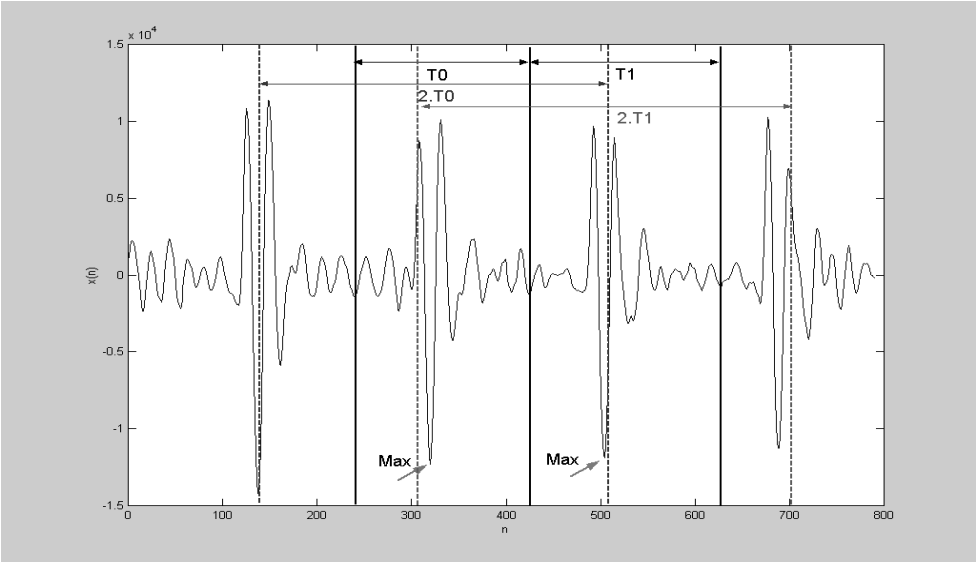

Fig. 3: *TD PSOLA – extraction of segments*

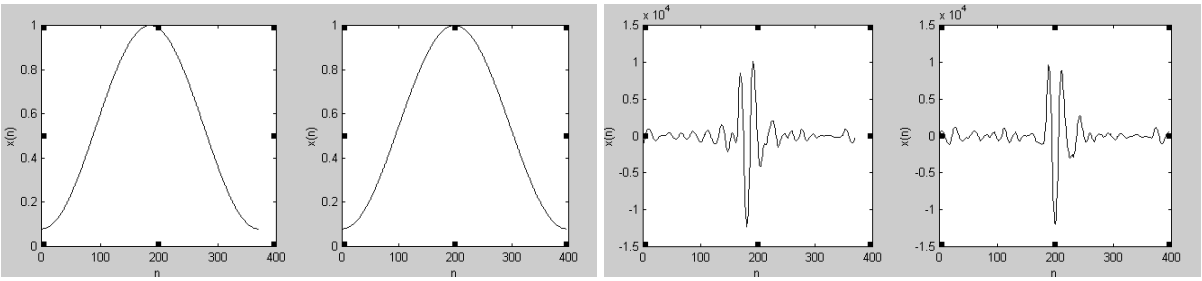

Fig. 4: *a,b) Hammings' windows corresponding to extracted signals; c,d) Extracted signals after windowing*

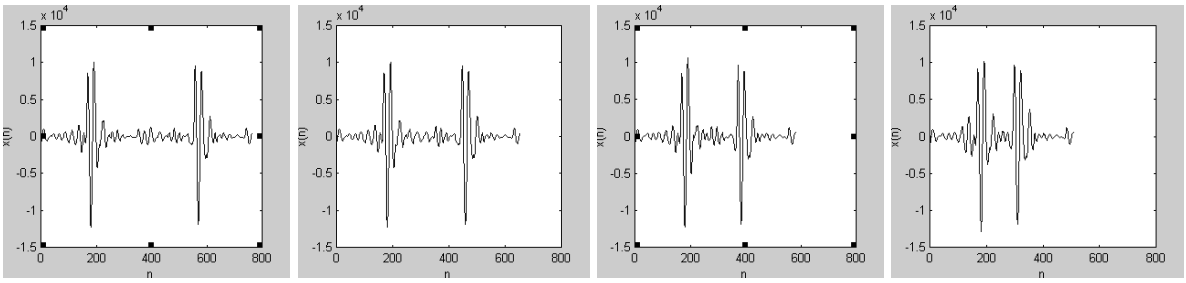

Fig. 5: *a) Summed signals without overlap; b,c,d) Signals with overlap from 0,7.T, 0,5.T, 0,3. T*

6 CONCLUSIONS AND PERPECTIVES

However, the speech, synthesized by TD PSOLA algorithm is quite understandable, human ear perceives it as synthetic. Furthermore, a great elongation of synthesized speech signal by adding periods, causes the echo effect. Contraction of the signal does not cause unfavourable auditive perceptions. However, omitting more periods in diphone, causes the loss of speech information content.

The TD PSOLA algorithm was implemented in TCL scripting language. As a result of belonging to the group of interpreted languages, TCL is relatively slow. Even though the programme was optimized, the synthesis of a sentence lasts several seconds. I suppose, that the use of compiled language e.g. C++, combined with hardware performance growth in the future, could shorten this time to hundreds of miliseconds. The undeniable advantage of TCL language remains the possibility of porting TCL scripts among operating systems.

The fundamental problem of the EMU TCL console is a plenty of errors in source code. Many functions does not work properly, and some of them, in special cases, does not work at all. Therefore some of them had to be newly implemented.

Despite the fact that the speech created by concatenation of diphones is superior to concatenation of phonemes, this approach does not enable synthesis of high quality human speech. It is caused by the fact, that neither a huge diphone database is able to cover a great variety of human speech. The use of MultiBand Excitation Resynthesis on database, might slightly improve the quality of the speech, but articulate synthesizers remain the vision for the future.

REFERENCES

[1] Dutoit, T.: An Introduction to Text-to-speech Synthesis, http://tcts.fpms.ac.be/synthesis

[2] Cassidy, S., Harrington J.: Multi-level Annotation in the EMU Speech Database Management System, Speech Communication 33, 2002

[3] Dutoit, T., Leich, H.: MBR PSOLA TTS Synthesis Based on an MBE Re-Synthesis of the Segments Database, http://tcts.fpms.ac.be

[4] Syrdal, A. and Col.: TD PSOLA Versus Harmonic Noise Model In Diphone Based Speech Synthesis, http://www.zippy.ho.att.com