Text-to-Speech Systems for Romanian Language

Gavril TODEREAN, Ovidiu BUZA, András BALOGH
Technical University of Cluj-Napoca
T.Gavril@yahoo.com, Ovidiu.Buza@com.utcluj.ro, Andras.Balogh@com.utcluj.ro, Adriana.Stan@com.utcluj.ro

***Abstract* — This paper presents the text-to-speech systems developed by our team at the Technical University of Cluj-Napoca. The speech-synthesis methods we implemented are the formantic, the phoneme- and diphoneme-based, the Multipulse Excited LPC, the Nonlinear Springing Method, the Repeat-Remove PSOLA and the HMM-TTS called HTS.**

***Index Terms* — HTS, LightVox, Romanian, RomVox, SprintVox, text-to-speech**

I. INTRODUCTION

This paper presents our achievements in text-based voice synthesis for Romanian language. This work inspired six doctoral thesis, all finalised by the members of our research team at the Technical University of Cluj-Napoca (UTCN).

II. THE BEGINNING

Our work started in the early '90s, with a theoretical study of the analysis and synthesis of the speech signals [16], [17], followed by the implementation of several algorithms on diferent hardware platforms.

After improving the analysis and synthesis algorithms, a first tex-to-speech system was developed, based on a software simulation on IBM-PC [7]

This platform offered the possibility to conduct experiments on some prosody imposing algorithms and to further improve the automatic phonetic transcription for the Romanian language. These algorithms were used to build the first unrestricted TTS system for Romanian, commercially named ROMVOX [5], which was relaying on a proprietary database with more than 1300 mono-, di- and triphones of the spoken language.

The formantic synthesis was experimented with PHILIPS's specialized I2C circuit, the PCF 8200.

Next platform was the TMS320C25, on which a classical LPC, then the Multipulse-Excitation LPC were developed [6], the second one being included into the ROMVOX system.

III. THE SPRINTVOX SYNTHESIS SYSTEM

A series of research has been carried out to impose the desired prosody of a synthesized text. A first method was the *Nonlinear Springing Method* (NSM) [1].

This time-based method achieves the imposition of the desired prozody using a predicted prosody matrix and a processed waveform annotated with specific information on each fundamental period: fundamental frequency variation, the intensity contour, and the temporal structure.

To control the duration and frequency of sound, some signal periods must be either reinserted or omitted. In the first step, the number of periods for each elementary sound is calculated. After the concatenation of the required diphones, the number of periods for each sound is determined based on the fundamental period descriptors. In order to keep the number of the samples of the input signal, a resampling of the synthesized signal is required by imposing the intonation curve with the predicted intonation profile.

One of the advantages of this method is that there are no discontinuity points between the fundamental periods of the generated signal. The other major advantage is that imposing the intonation curve can be done with great precision.

Another direction of research to find a better method of imposing prosody has proposed to improve the TD-PSOLA method. The theoretical study in this field and the results obtained were presented in [3]. The resulting method from this study was called *Repeat-Remove Controlled PSOLA* (RR-PSOLA) [4]. For this synthesis technique, a database of about 1300 diphones ([2], [3]) was designed and realized. The database annotation was made automatically and includes the marking of the fundamental periods -the beginning and the end of each period, the type of the period -marked as "voiced", "unvoiced" or "complex", the repeatability or removability of a period and phonemic limits within diphones.

The advantages of this method are as follows:
- by fixing the fundamental periods that are not situated on the stable parts of the phonemes, a more efficient spectral envelope and a much better amplitude of the resulting signal can be generated;
- the method keeps intact the transitions between the phonemes within a diphone, thus preserving as much as possible the naturalness of the acoustic units in the synthesized speech;
- marking the limits of phonemes in a diphone can separate the prosody parameters of each phoneme, resulting in the advantage of more accurate prosody generation.

Based on the RR-PSOLA voice synthesis method, a TTS system for Romanian language called SprintVox was designed and built [2],[4]. The system architecture (Fig. 1) comprises three levels: the command level, the operations level, and the database level.

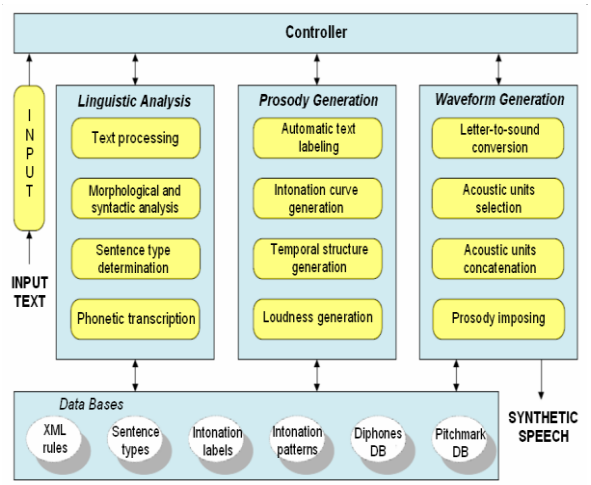

Fig. 1. The SprintVox System Architecture

A. The command level is responsible for controlling the running sequence of component modules and for transferring data between modules.

B. The operations module level is grouped into three sections: linguistic analysis, prozody generation and waveform generation.

a) Linguistic analysis performs text preprocessing, morphological, syntactic and semantic analysis, sentence type decision, phonetic transcription, and automatic text labeling.

Determining the type of sentences in the input text makes it possible to select a particular pattern of intonation that will be applied to the generated waveform. In order to identify the types of sentences in Romanian, a study of the intonation patterns was made and the basic types of sentences from the Romanian language were determined and a database of these types was made [4]. The types of sentences are characterized by a certain intonation pattern and are classified based on their morphological structure and certain keywords that they contain. In [3] are presented the basic types that were determined for the Romanian language, as well as the classification rules.

b) Then, the prozody generation assumes the realization of the intonation curve, the generation of the temporal structure and the generation of sound intensity.

c) The waveform generation consists of the literal-sound conversion, the selection of the acoustic units, unit concatenation and the prozody imposition by the RR-PSOLA method.

C. The database level contains the parametric acoustic units, the intonation curve patterns, and XML rules for synthesized text. The database was specially designed to be easily expandable, by modifying the files containing the processing rules [2],[3].

In the final implementation, the Sprintvox system was incorporated into the Audacity open source program.

IV. THE LIGHTVOX SYNTHESIS SYSTEM

Another contribution in the field of voice synthesis was the realization of a synthesis system adapted to the Romanian language, using syllables as phonetic units. The system, called LIGHTVOX [8], [9], was conceived as a text-to-speech system, in which the speech synthesis is done by an original method, starting from the analysis of a text in Romanian language [9].

The LIGHTVOX system (Fig. 2) is organized into five component modules: the linguistic analysis module, the prosody analysis module, the vocal database management module, the phonetic units matching module, and the speech synthesis module.

The linguistic analysis module performs the analysis of the input text in order to extract the basic phonetic units, namely the syllables. The syllable based design was chosen because the Romanian language has a specific rhythm of speech in which syllables are easy to detect. The use of syllables for synthesis also leads to more advantages: increases the voice naturalness, easy maintenance of the vocal database due to the relatively small number of syllables in Romanian, less concatenation errors due to the reduced number of interpolation points inside a word [9].

The prosodic analysis module identifies segmental prosodical elements based on the input text. In the first phase, the places of speech accentuation within words are determined, based on a set of rules specifically designed for Romanian [9].

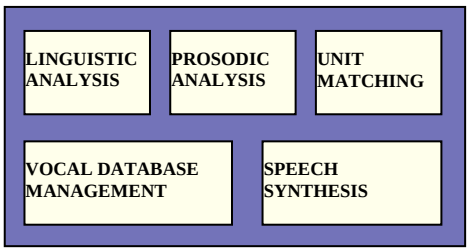

Fig. 2. The LightVox System

The vocal database management module performs all operations related to the acoustic units database. The vocal database includes a subset of Romanian syllables obtained from recording of a speaker utterances. The database is organized according to the type of syllables, their position within the word, as well as the number of component phonemes [10].

The phonetic units matching module realizes the match between the syllables extracted from the input text and the syllables recorded in the vocal database in acoustic format as waveforms. The matching process must be done in an optimal manner, taking into account that not all Romanian syllables are stored in the database.

The speech synthesis module concatenates the waveforms provided by the phonetic units matching module. The output sound is obtained by calling API functions that control the computer's audio card.

A. In the first step of text analysis, there are identified the linguistic units: sentences, words, and syllables. This is done through a lexical analyzer based on linguistic rules. The same lexical analyzer is also used to detect separator characters, special characters and numbers. The analyzer is

made using the LEX parser generator, which generates a text analyzer based on a grammar that describes the rules for parsing the text. The grammar is written in the standard BNF format, specifying the sequences of characters that can be recognized from the input text. These sequences refer to syllables, separator characters, special characters and numbers [9],[10].

B. In the second stage of analysis, the prosodic information is determined, this meaning in our case to determine the position of the accent within the words. This phase is accomplished through a phonetic analyzer, which takes each word from the text and chooses the accentuated syllable based on phonetic rules. Phonetic rules are also written in standard BNF format. In Romanian, the accent may fall on any of the last four syllables of the word, but most often on the penultimate syllable. The set of phonetic rules contains this basic rule, plus a consistent set of exceptions, organized into classes of words having the same termination [9],[10].

C. The vocal database includes Romanian syllables, recorded as PCM waveforms. So far, only a subset of the syllables of the Romanian language has been registered. Syllable units have been inserted into a tree structure following this classification:

- after syllable length (two, three or four phonemes);
- after the syllable position within the word (beginning, median or final);
- after accentuation (accentuated or non-accented syllables).

The matching of the syllables identified from the input text with the acoustic units recorded in the vocal database is done accordingly with this units classification. If a syllable is not found in the database, it will be constructed from other existing syllables and phonemes. The vocal database currently contains 562 syllables.

V. SISTEMUL DE SINTEZĂ DE VOCE HTS

The research work conducted in order to obtain a corpus-based synthesis for romanian [13] resulted in a high quality TTS system, the HTS (*Hidden Markov Model based Text-to-Speech System*).

Within this work a 65000-entry phonetic dictionary was created, containing the phonetic transcription and the accent's position for each pre-recorded word. The transcription was carried out with an automatic method, using the standard phonemes-set of the Romanian language. The basic rules for the letter-to-sound conversion are written in the Festival Speech Synthesis System (free software), then the Romanian-specific rules were added manually to the dictionary. The accents were extracted directly from the DEX on-line database [13], [15].

The realized dictionary –because it's size and quality- and the method developed for it's further extension are valuable linguistic resources. The correctness of the database content was tested automatically, being used for training the HTS system.

A voice-corpus for Romanian, called RSS (*Romanian Speech Synthesis*) was also recorded [13], containing 3.5 hours of recorded speech signal. This was used for training the system and another set of 0.5 hours of recordings for it's testing. Besides audio recordings, the RSS contains the phonetical and orthographical transcriptions of the phrases, HMS labels for speech synthesis and the accent's positions inside words. (Fig. 3).

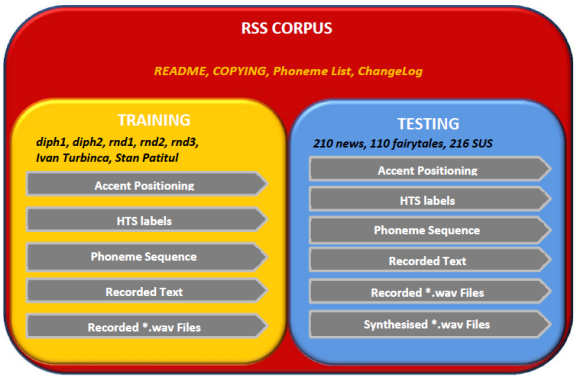

Fig. 3. The graphical user interface of the RSS

The HTS development was realized in a dynamic way, evaluating many of it's possible configuration options, trying different spectral analysis methods, different ways to compute the MFCC (*Mel Frequency Cepstral Coefficients*), different sampling rates for the recorded units and the size of the training set.

During the HTS development a new modelling technique of the fundamental frequency was designed. This is based on the Discrete Cosine Transform [14] and also an interactive method for prosody-adjustment [15]. Both methods are coming to enhance the overall system performance in terms of similarity between the voice of the original speaker and the synthetized voice and in terms of naturalness on the MOS scale, reducing the rate of erronous output words. These results for the HTS quality evaluation are exemplified in Fig. 4, being detailed în [15].

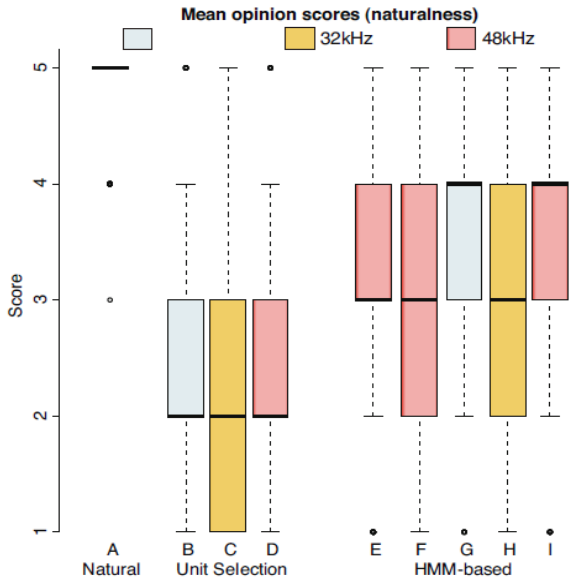

Fig. 4. The naturalness evaluation of HTS

VI. CONCLUSION

This paper presents the text-to-speech systems and voice synthesis methods realized by our research team at UTCN, all these systems being an important step forward in the field of voice synthesis for Romanian language.

REFERENCES

[1] Á. Zs. Bodó, G. Toderean, *Experiments for prosody modification using the Nonlinear Springing Method*, Trends in Speech Technology, SpeD 2005, Cluj-Napoca, pp. 177-181,.

[2] Á. Zs. Bodó, O. Buza, G. Toderean, A*coustic Database for Romanian TTS Synthesis. Design and Realisation Results (I)*, Acta Technica Napocensis, Vol 48, No. 2/2007, Cluj-Napoca, pp. 24-31

[3] Á. Zs. Bodó, *Contribuţii la sinteza vorbirii în limba română*, teză de doctorat, Universitatea Tehnică din Cluj-Napoca, conducător ştiinţific prof. G. Toderean.

[4] Á. Zs. Bodó, O. Buza, G. Toderean, *Experiments with the prediction and generation of Romanian intonation*, "Speech Technology and Human-Computer Dialogue", SpeD 2009, Constanţa

[5] Ferencz A., D. Zaiu, T. Ratiu, M. Ferencz, Toderean G. *Speech Synthesis from Unrestricted Text & the ROMVOX System for Romanian Language*, Revista STUDIA sem.IV.

[6] Ferencz A., Arsinte R., Raţiu T., Ferencz M., Zaiu D., Toderean G., *Experimental Implementation of the LPC-MPE (multi-pulse excitation) Synthesis Method for the ROMVOX Text-to-Speech System*, Speech and Computer (SPECOM'96) International Workshop, St. Petersburg, Rusia, pp. 159-154.

[7] Ferencz A., *Dezvoltare hardware-software pentru sinteza "Text-Vorbire" în limba română pe calculatoare compatibile IBM-PC*, teză de doctorat, Universitatea Tehnică din Cluj-Napoca, conducător ştiinţific prof. G. Toderean.

[8] O. Buza, G. Toderean, A. Nica, Zs. Bodo, *Original Method for Romanian Text-to-Speech Synthesis Based on Syllable Concatenation*, the Proceedings of the 4-th Conference on Speech Technology and Human Computer Dialogue "SpeD 2007", published in the volume "Advances in Spoken Language Technology", Iasi, Romania, pp. 109-118.

[9] O. Buza, G. Toderean, J. Domokos, A. Zs. Bodo, *Building a Text to Speech System for Romanian through Concatenation*, The 5th IEEE Conference on Speech Technology and Human Computer Dialogue SpeD 2009, Constanta, Romania.

[10] O. Buza, *Contribuţii la analiza şi sinteza vorbirii din text pentru limba română*, teză de doctorat, Universitatea Tehnică din Cluj-Napoca, conducător ştiinţific prof. G. Toderean

[11] O. Buza, G. Toderean, J. Domokos, *Automatic Algorithm for Region Segmentation of Speech Signal*, The 4-th International Conference on Communications, Mobility, and Computing (CMC 2012), Guilin, China., pp.179-182.

[12] O. Buza, Toderean G., Balogh A., Domokos, J., *Algorithm for detection of voice signal periodicity*, the 7th Conference on Speech Technology and Human - Computer Dialogue (SpeD 2013), Cluj-Napoca, Romania

[13] A. Stan, ş.a., *The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate*, Speech Communication, vol 53, pg. 442-450.

[14] A. Stan, Adriana Stan, F.-C. Pop, M. Cremene, M.Giurgiu, D.Pallez, *Interactive Intonation Optimisation Using CMA-ES and DCT Parametrisation of the F0 Contour for Speech Synthesis*, In Proceedings of the 5th Workshop on Nature Inspired Cooperative Strategies for Optimisation, in series Studies in Computational Intelligence, vol. 387, Springer, Berlin.

[15] A. Stan, *Sinteza text-vorbire în limba română bazată pe modele Markov şi optimizarea interactivă a intonaţiei*, teză de doctorat, Universitatea Tehnică din Cluj-Napoca, conducător ştiinţific prof. Mircea Giurgiu.

[16] G. Toderean, A. Căruntu, *Metode de Recunoaştere a Vorbirii*, Editura Risoprint, Cluj-Napoca

[17] G. Toderean, O. Buza, Á. Zs. Bodó, *Metode de Sinteză a Vorbirii*, Editura Risoprint, Cluj-Napoca.