A Rule-Based Approach to Build a Text-to-Speech System for Romanian

Ovidiu Buza, Gavril Toderean, Jozsef Domokos
Faculty of Electronics and Telecommunications
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
Ovidiu.Buza@com.utcluj.ro, toderean@pro3soft.ro, jdomokos@com.utcluj.ro

Abstract—We present in this article our approach for building a text-to-speech system for Romanian. Main stages of this work were: voice signal analysis, region segmentation, construction of acoustic database, text analysis, unit and prosody detection, unit matching, concatenation and speech synthesis. In our approach we consider word syllables as basic units and stress indicating intrasegmental prosody. A special characteristic of current approach is rule-based processing of both speech signal analyse and text analyse stages.

Keywords—text-to-speech; rule-based approach; syllable detection

I. INTRODUCTION

This article presents our experience in building a voice synthesis system complying with quality parameters of human speech. Our researches led into projecting a voice synthesis method specifically adapted to Romanian language, and also into a working approach for constructing an automated speech synthesis system.

Using syllables as basic units, the projected method is integrated into high quality methods category, based on concatenation. The novelty of this work consists in the fact that our approach is based on rules that apply in the most important stages of speech synthesis system realisation (Fig. 1):

a) signal processing stage, by extracting acoustic units from speech using association rules,

b) text processing stage, by extracting linguistic units using phonetic and lexical rules.

![Signal Regions](images/signal_regions.png)
![Text](images/text.png)

In signal processing stage, we have extracted main parameters of pre-recorded speech, parameters that will be used in speech segmentation. We have designed an algorithm for signal decomposition in classes of regions, that are associated with phonetic categories of Romanian language. Then we have used a semi-automated algorithm for separating syllables from speech signal and storing them into vocal database.

In text processing stage, special phonetic rules have been developed for text processing, linguistic units (syllables) detection, and prosody data (like stress) retrieval.

Next, unit matching was done by selecting acoustic units from vocal database according to the linguistic units detected from the input text. And finally, acoustic units are concatenated to form the output speech signal, that is synthesized by the mean of a digital audio card.

II. SIGNAL PROCESSING

Voice signal processing starts with the detection of signal parameters from recorded speech. This process can be done in time or frequency domain. Time domain processing, that we have used, leads to the detection of signal parameters directly from waveform samples. We have extracted following parameters: maximum and median amplitude, signal energy, number of zero-crosses and fundamental frequency.

Signal amplitude [7] gives information about presence or absence of speech, about voiced and unvoiced features of the signal on analyzed segment. In the case of a voiced segment of speech, as a vowel utterance, the amplitude is higher, beside the case of an unvoiced speech segment, where amplitude is lower.

Signal energy [8] is used for getting the characteristics of transported power of speech signal. Voiced segments (like vowels) have a higher mean energy, while the unvoiced segments (like fricative consonants) have a lower mean energy. For the majority speech segments, energy is concentrated in 300-3000 Hz band.

Number of zero-crosses is used for determining frequency characteristics and voiced/unvoiced features of speech segments [7]. Inside voiced segments number of zero-crosses is higher, while inside unvoiced segments this parameter has much lower values.
III. SPEECH SEGMENTATION

Finding an optimal approach for speech signal segmentation is an imperative in building acoustic database of a voice synthesis system. This section presents a segmentation method that has been designed and implemented by the authors, method which is capable to detect S/U/V (Silence, Unvoiced, Voiced) components of speech signal, to divide these components in regions with specific characteristics and to associate regions with a known phonetic input sequence.

A. S/U/V Segmentation

Our segmentation method uses time domain analysis of speech signal. After low-pass filtering of signal, zero-crossing waveform points (Z) are detected. Then minimum ($m_i$) and maximum ($M_i$) values between two adjacent zero points are computed.

Separation between silence and speech segments is realised by using a threshold value $T_s$ on signal amplitude. In silence segments, $m_i$ and $M_i$ points must be lower than threshold $T_s$.

For speech segments, distance $D_i$ between two adjacent zero points is computed. Decision of voiced segment is assumed if distance $D_i$ is greater than a threshold distance $F$. A segment is assumed unvoiced if distance between two adjacent zeros is smaller than a threshold $U$. Transient segments are also defined and they consist of regions for which above conditions are not accomplished. $V$ and $U$ thresholds have been chosen according to statistical median frequency for vowels and fricative consonants.

After first appliance of above algorithm, a large set of regions will be created. Since voiced regions are well determined, the unvoiced are broken by intercalated silence regions. This situation appears because unvoiced consonants have low amplitude so they can break in many silence/unvoiced subregions.

All these regions will be packed together in the second pass of the algorithm, so the result will be a single unvoiced region.

After segmentation, voiced and unvoiced segments are coupled according to the syllable chain that is used in vocal database construction process. Appropriate acoustic units will be detected, labeled and stored in vocal database.

B. Regions detection

The SUV segmentation process presented above divides the signal in four basic categories: Silence, Voiced, Unvoiced and Transition. Each category will be further classified in distinct types of regions, totally 10 classes: silence, unvoiced-silence, voiced, voiced glide, voiced plosive, voiced jump, transition, irregular (rugged), high density and unvoiced consonant (Fig. 2). The aim of this classification is to associate Romanian phonemes with signal regions having some particular traits.

These ten classes of regions are:

1. Silence region (S). Represents a region without speech, where signal amplitude is very low.
2. Unvoiced Silence (US). This region is a combination of silence S and unvoiced consonant C region.
3. Voiced Region (V). The voiced region contains all Romanian vowels: /A/, /E/, /I/, /O/, /Î/, /Â/, /Â/ glide /L/, nasals /M/, /N/, and voiced plosive consonants as /P/, /B/, /D/.
4. Voiced Glide (VG). This is a region corresponding to a glide consonant like /R/.
5. Voiced Plosive (VP). A region of voiced discontinuity associated with plosive consonants like /C/ or /G/.
6. Voiced Jump (VJ). Is a region similar with voiced region V, but it has no periodicity.
7. Irregular Region (IR). This is a region in which one can find plosive consonants like /C/, /G/ or /P/.
8. High Density transition (HD). Is a transition region with high frequency values, which could indicate emergency of fricative consonants.
9. Transition (T). Is an intermediate region between voiced and unvoiced and which has no the characteristics of IR or HD classes.
10. Unvoiced Consonant (C). For Romanian language, this class contains fricative consonants /S/, /S/, /Ţ/, /F/, /Z/, /H/, and non- fricatives /Ce/, /Cl/, /Ge/, /Gi/.

For detection of all these classes, median amplitude, number of zero-cross points, signal energy have been used, and also short-term Fourier coefficients for detection of VP and HD special cases.

C. Phonetic segmentation

Phonetic segmentation is the process of associating phonetic symbols with the speech signal. This process is very useful when we want to develop an acoustic database from a large speech corpus. Phonetic segmentation gives the capability of detecting and separating phonetic units from speech, units that will be used in achieving the output acoustic chain sequence through concatenation.

Phonetic segmentation method proposed by the authors is based on special association rules, which realise a correspondence between phonetic groups taken from an input stream and distinct types of regions detected from the speech signal. Segmentation algorithm parses the input text and tries to find the best match for each phonetic group with one or more regions of speech signal.

Input text is first written in phonetic transcription, using a simple look-after table, which includes phoneme and word transitions. Transcribed text is splitted into a sequence of phonetic groups. Special association rules will establish a correspondence with specific regions detected from speech signal.
For associating phonetic groups with sequences of regions, as further in our approach, we have used LEX parser generator [4]. We have written a set of association rules, each rule specifying a phonetic pattern for associating a particular group with a sequence of regions, and also specifying a condition to be verified in order to make that association. Each condition outlines: type of region, minimal and maximal duration, type of association: unique region or sequence of regions.

IV. BUILDING THE VOCAL DATABASE

Phonetic segmentation method described in previous section has been designed for segmentation and labelling of speech corpora, having as main objective the construction of vocal database of our speech synthesis system. In our approach, realisation of vocal database implies separation of acoustic segments that correspond with phonetic syllables of Romanian language and storing these segments into a hierarchical structure.

The speech corpus used for extracting acoustic units was built from common Romanian sentences, from separate words containing syllables, and also from artificial words constructed for the emphasis of one specific syllable. After recording, speech signal was normalized in pitch and amplitude. Then phonetic segmentation was applied and acoustic syllables were stored in database.

Units are stored in database following this classification:
- after length of syllables: we have two, three or four characters syllables (denoted S2, S3 and S4) and also singular phonemes (S1);
- after position inside the word: initial or median (M) and final syllables (F);
- after accentuation: stressed (A) or normal (N) syllables.

V. TEXT PROCESSING

Text processing stage implies realisation of following tasks:
- a) detection of linguistic units: sentences, words and syllables;
- b) generating prosody information, i.e. stress position within words.

Text processing tasks are accomplished by four modules that have been designed for unit detection, prosody data retrieval and unit processing. These modules are:
- a lexical analyzer for detection of basic units;
- a phonetic analyzer for generating prosody information;
- a high level analyzer for detection of high-level units;
- processing shell for unit processing.

Lexical analyzer extracts text characters and clusters them into basic units. We refer to the detection of alphabetical characters, numerical characters, special characters and punctuation marks. Using special lexical rules (that have been presented in [9] - [12]), alphabetical characters are clustered as syllables, digits are clustered as numbers and special characters and punctuation marks are used in determining of word and sentence boundaries.

Phonetic analyzer gets the syllables between two breaking characters and detects stress position, i.e. the accentuated syllable from corresponding word.

Then, high-level analyzer takes the syllables, special characters and numbers provided by the lexical analyzer, and also prosodic information, and constructs high-level units: words and sentences. Also basic sentence verification is done here.

Processing shell finally takes linguistic units provided from the previous levels, classifies and stores them in appropriate structures. From these structures, synthesis module will construct the acoustic waves and will synthesize the text.

A. Syllable detection

A lexical analyzer has been designed for detection of basic units: syllables, breaking characters and numbers. This module was constructed by using LEX scanner generator [4]. LEX generates a lexical scanner starting from an input grammar that describes parsing rules. Grammar is written in BNF standard form and specifies character sequences that can be recognized from the input. These sequences refer to syllables, special characters, separators and numbers. For detection of syllables we have been inspired from Romanian syntax rules [3].

Hereby, input text is interpreted as a character string. At the beginning, current character is classified in following categories: digit, alphanumeric character, and special character. Taking into account left and right context, current character and the characters already parsed are grouped to form a lexical unit: a syllable, a number or a separator. Specific production rules for each category indicate the mode each lexical unit is formed and classified.

Regarding rule matching process inside lexical analyzer, two types of rule sets were made: a basic set consisting of three general rules, and a large set of exception rules which states the exceptions from the basic set.

Basic set shows the general decomposition rules for Romanian:

\[
\begin{align*}
\text{syllable} &= \text{CONS}\{\text{VOC}\}^* \text{CONS} \quad (R1) \\
\text{syllable} &= \text{CONS}\{\text{VOC}\}^*\{\text{CONS}\}/\{\text{CONS}\} \quad (R2) \\
\text{syllable} &= \text{CONS}\{\text{VOC}\}^*\{\text{CONS}\}/\{\text{SEP}\} \quad (R3)
\end{align*}
\]

First rule (R1) assumes that a syllable can be a sequence of consonants followed by a vowel. Second rule (R2) states that a syllable can be finished by a consonant if the beginning of the next syllable is also a consonant. Third rule (R3) says that one or more consonants can be placed at the final part of a syllable if this is the last syllable of a word.

Exception set is made up from the rules that are exceptions from the three rules of above. These exceptions are situated in the front of basic rules. If no rule from the exception set is matched, then the syllable is treated by the basic rules. At this time, exception set is made up by more then 180 rules. Rules are grouped in subsets that refer to resembling character sequences. All these rules were completely explained in [9].
B. Syllable accentuation

The principle for determining syllable accentuation resembles with that of lexical analyzer for detecting syllables already exposed. After the text parser returns from input stream current word consisting of phonemes $F_1, F_2, ..., F_k$ and delimited by a separator, phonetic analyzer reads this word and detects syllable accentuation based on phonetic rules. Rules have been also written in BNF form and set into LEX input.

In Romanian, stressed syllable can be one of last four syllables of the word: $S_n, S_{n-1}, S_{n-2}$ or $S_{n-3}$ ($S_n$ is the last syllable). Most often, stress is placed at next to last position.

Rules set for determining accentuation consists of:

a) One general rule meaning $S_{n-1}$ syllable is stressed:

$$\{\text{LIT}\}+/{\text{SEP}} \{ \text{return(SN-1);} \} \quad (G1)$$

b) A consistent set of exceptions, organized in classes of words having the same termination. At this time, the exception set is made up by more than 250 rules. All these rules have been presented in [9].

VI. UNIT MATCHING, CONCATENATION AND SYNTHESIS

Matching process is done according to the three-layer classification of units: number of characters in the syllable, accentuation and the place of syllable inside the word. If one syllable is not found in vocal database, this will be constructed from other syllables and separate phonemes that are also recorded. Following situations may appear:

(a) Syllable is matched in appropriate accentuated form. In this case acoustic unit will be directly used for concatenation.

(b) Syllable is matched but not the accentuation. In this case, unit is reconstructed from other syllables and phonemes which abide by the necessary accentuation.

(c) Syllable is not matched at all, so it will be constructed from separate phonemes.

After matching, units are simply concatenated to result the acoustic chain that will be synthesized. In this stage of development, our system works with intrasegmental prosody, i.e., accentuation inside words, and doesn’t support sentence-level prosody like intonation. Rhythm of speech can be adjusted by intercalating different periods of silence between syllables, words and sentences.

VII. CONCLUSIONS AND RESULTS

We have presented in this article a complete method for building a syllable-based text-to-speech system. First, speech signal was segmented in basic categories and ten different classes of regions. Then, a rule-based segmentation method was invoked onto a parallel text&speech corpus in order to associate input phonemes with regions. We have used this segmentation method to separate acoustic units from speech corpora and create the vocal database. Performance evaluation of automated segmentation method shows a segmentation error of maximum 5% comparative to a manual approach.

Special efforts have been done to accomplish the text processing stage. Here we have designed two sets of rules: one set of rules for detecting word syllables and a second set for determining the accentuation inside each word. Even these sets are not complete, they cover yet a good majority of cases. The lexical analyzer is based on rules that assure more than 98% correct syllable detection, since accentuation analyzer provides about 94% correct detection rate (computed on near 50000 words database consisting of different Romanian texts).

The advantages of detecting syllables through a rules-driven analyzer are: separation between syllables detection and system code, facile readability and accessibility of rules. Other authors [1] have used LEX only for pre-processing stage of text analysis, and not for units detection process itself. Some methods support only a restricted domain [6], since our method supports all Romanian vocabulary. The rules-driven method also needs fewer resources than dictionary-based methods [5].

About speech synthesis outcome, the results are encouraging, and after a post-recording stage of syllable normalization we have obtained a good, near-natural quality of speech synthesis. For the future implementations, we have in mind the completion of syllable and accentuation rules sets and also the completion of syllables database according to a self-generated statistics of Romanian syllables, aspects that will improve the system performance.

REFERENCES


