The principles of designing of algorithm for speech synthesis from texts written in Albanian language

Agni Dika[1], Adnan Maxhuni[1], Avni Rexhepi[1]

[1] Faculty of Electrical and Computer Engineering, University of Pristina, Kosovo

Abstract

The speech synthesis is artificial generation of human speech from written texts. For this purpose, adequate algorithms are designed, which then through relevant programs make it possible to synthesize texts to speech. The process of converting text into speech is also known as Text-To-Speech (TTS) system [5].

In this paper are given basic principles to be used when designing a system to synthesize speech in Albanian language[1] from written texts. Currently there are solutions that enable natural speech generation for various world languages. However, unfortunately these are not universal solutions to be used for other languages too, because the volume generated for other languages is incomprehensible and unnatural. For this reason, for every language one should seek solutions that address the specifics of it, always with the aim of generating voice to suit the nature of language. Generating systems that are currently used mainly rely on the use of the concatenation method [6], during which acoustic segments of text files are joined, which are previously digitized and stored as such in a database.

For Albanian language, we consider that on the textual part of the database, as basic segments to be used are: the most frequent words, two-letters and letters [4]. However, in a particular part of the database are included various abbreviations, i.e. textual equivalents and their acoustics files, to be used also during the generation of appropriate speech. Whereas, with the aim of synthesizing the various numerical values written in the decimal system, in database were added values, respectively their corresponding sound files, whereby speech is generated for different numbers. The first part of the paper is a brief presentation of the Albanian language [1], respectively of the alphabet used in writing the language and its most frequent words.

Keywords: *TTS, Albanian, Concatenation, Synthesis*

1. Introduction

Among people there are some that are unable to read, either because of blindness (complete or partial), or for other reasons. Therefore, initially for this category of people, there is a need for computer-generated speech, namely the conversion of written texts into acoustic files in Albanian language [7].

In the market, there already are different solutions to synthesize speech from text written in different languages. However, these solutions cannot be used for generating speech in other languages, because for every language, specific algorithms should be used to synthesize speech, since each language is different during the speech. Therefore, for the Albanian language too, there is a need to design a particular algorithm and to create the system to synthesize speech in this language.

From the published works, it is known that to synthesize speech from written text, generally three different methods are used: concatenative synthesis, formant synthesis and articulatory synthesis. Each of these methods has advantages and disadvantages. Quality of synthesis is assessed by *naturality* and

Based on the project financed by Ministry of Education, Science and Technology of Kosova

comprehensibility of the generated speech [6]. Naturality has to do with the likeness of the generated speech to the human speech, while, comprehensibility has to do with the clarity of the generated speech.

2. Albanian language features associated with synthesis of speech

While designing the system for conversion of written text into speech, certainly of particular importance are the characteristics of the language which is synthesized. For the purpose of emphasis of basic characteristics of the Albanian language, follows a short presentation of it.

2.1. Frequency of letters

The Albanian language belongs to Indo-European family of languages [1], which also includes languages as Greek, Roman, German, etc. In the tree of languages, Albanian language is presented with a special branch, which comes somewhere from the roots of this tree. It has common features with these languages, but it also has differences, which characterize it.

The Albanian language is written using the alphabet with 36 letters (Table 1), which is created based on the Latin alphabet.

Table 1. Albanian alphabet

A a	B b	C c	Ç ç	D d	Dh dh
E e	Ë ë	F f	G g	Gj gj	H h
I i	J j	K k	L l	Ll ll	M m
N n	Nj nj	O o	P p	Q q	R r
Rr rr	S s	Sh sh	T t	Th th	U u
V v	Y y	X x	Xh xh	Z z	Zh zh

Letters belong to 36 phonemes of the spoken Albanian language, so that for every phoneme that is generated during the speech, in the written texts the respective graphemes are used. Of these, 7 are vowels (a, e, ë, i, o, u and y), and 29 are consonants. In the group of consonants there are 9 two-symbol letters (dh, gj, ll, nj, rr, sh, th, xh and zh), which are known as "bigrames". Within the group of one-symbol letters, 25 of them are Latin Alphabet, while 2 are excluded and are known as diacritical characters - letters "ç" and "ë" (see Tab. 1). Two-symbol letters (bigrames) present a challenge in itself during the conversion of text to speech, because they must be treated as single letter.

Based on measurements, the frequency of use for the letters of the Albanian language is found, as given in Figure 1.

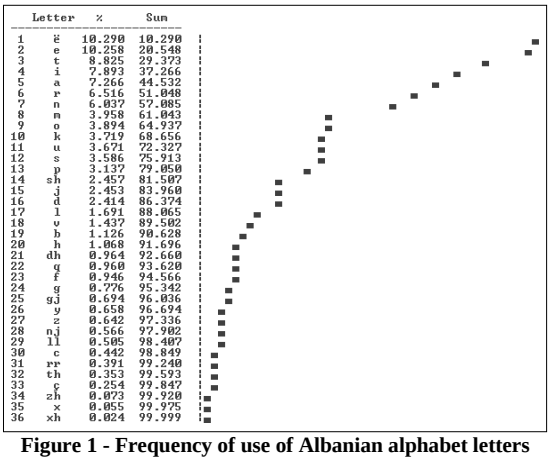
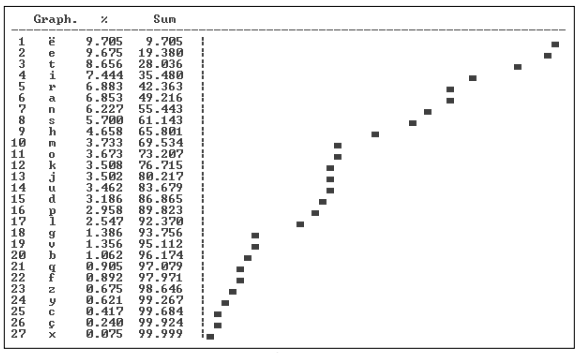


Figure 1 - Frequency of use of Albanian alphabet letters

The diagram shows that the greater frequency of use have vowels "ë" and "e". But, on the third place appears consonant "t", which is part of the word "të", which is calculated as the most frequent word in Albanian language. After that, as in many languages, as letters with high frequency are placed two other vowels: i and o.

In order to compare the frequency of use of Albanian language graphemes with those in other languages, in figure 2 is shown the distribution of frequencies of the use of graphemes, in Albanian.



	Graph.	%	Sum
1	ë	9.705	9.705
2	e	9.675	19.380
3	t	8.656	28.036
4	i	7.444	35.480
5	r	6.883	42.363
6	a	6.853	49.216
7	n	6.227	55.443
8	s	5.700	61.143
9	h	4.658	65.801
10	m	3.733	69.534
11	o	3.673	73.207
12	k	3.508	76.715
13	j	3.502	80.217
14	u	3.462	83.679
15	d	3.186	86.865
16	p	2.958	89.823
17	l	2.547	92.370
18	g	1.386	93.756
19	v	1.356	95.112
20	b	1.062	96.174
21	q	0.905	97.079
22	f	0.892	97.971
23	z	0.675	98.646
24	y	0.621	99.267
25	c	0.417	99.684
26	ç	0.240	99.924
27	x	0.075	99.999

Figure 2 - Distribution of graphemes, in texts written in Albanian

One of the common features of the Albanian language with other languages is the distribution of use of graphemes in the text. Thus, for example comparison of the graphemes' frequency distribution in English [2] and in Albanian language is shown in diagram given in figure 3.

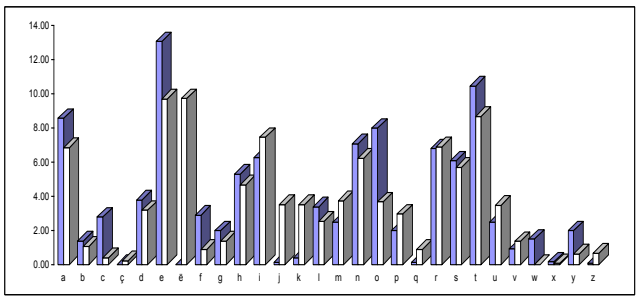


Figure 3 – Comparative diagram of the distribution of graphemes in Albanian and English

In the diagram, with dark colored columns are presented graphemes use frequencies in English. Simultaneously, each column is attached to the white column, belonging to the frequency of grapheme use in Albanian. Thus, for example, the first two columns in the diagram present use frequencies for grapheme "**a**" in English (**8.56%**) and Albanian (**6.85%**).

Despite the similarity of frequency of use of graphemes with English, Albanian language is unique in the way of creating, writing and reading words. For these reasons, after testing various applications provided for converting speech to text in English, applying Albanian texts, is proven lack of naturality and comprehensibility of speech created in Albanian. A slightly better approximation is seen with the generation of speech with synthesis through the use of Italian language synthesizer, but even this case lacks the natural and completely comprehensible generation.

A peculiarity of the Albanian language, compared to major world languages (e.g. English) is read of graphemes used as symbols of writing. This peculiarity relates to a unique reading of graphemes, regardless of where they are located. So, every letter is read in a unique way. This represents an advantage in case of conversion of text to speech. Naturally, this creates ease in reading specific words. Exception are 9 bigrams (two-symbol letters) written using two graphemes, and read as a single letter. To eliminate the problem of two-symbol letters, before generating sound files, they can be replaced with one-symbol, which makes it possible to treat them as ordinary letters.

An issue in themselves constitute two diacritics of Albanian language that cannot be generated directly from the computer keyboard. They are graphemes "**ë**" and "**ç**". While grapheme "**ç**" is very close to "q", grapheme "**ë**" is not similar to any other grapheme in reading. Moreover, it is a letter with very high frequency of use (almost the highest, as well as letter **e**).

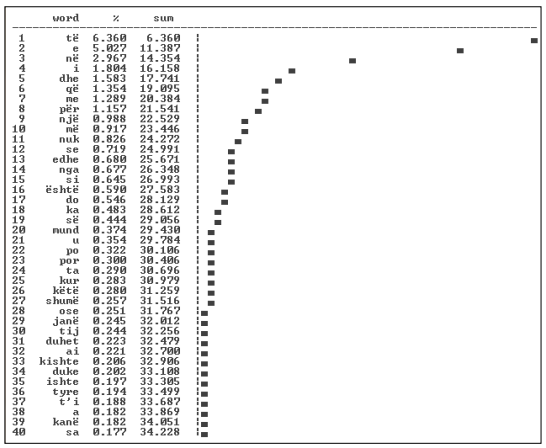
Another important issue in converting written text into speech, represent the numbers. Fortunately, in Albanian, most of the numbers are pronounced the same, regardless of whether they represent: the year, the numerical value, exchange value or other dimensions. Some of the exemptions (not many) are telephone numbers, address numbers, and any other occasion. Even this is a mitigating circumstance, compared to many other languages, like English, where

the reading of a number that represents the year, differs from reading/pronouncing a written number that represents a numeric value or exchange value. The idea is *firstly to convert numbers to text/words and afterwards the speech generation procedure applies as well as to the usual text*. A great convenience in this case is the fact that only a few words are used for "converting" numbers to words, i.e. the number of relevant sound files is small.

Another problem to be solved during the conversion of text in Albanian are spoken acronyms (abbreviations), e.g.: UN, UNICEF, NATO, EU, etc.. In these cases, reading the letter by letter does not represent a large deformation, so therefore it can be considered as permissible. However, for the most common acronyms, their pronunciation forms (words) can be stored, respectively, their corresponding sound files.

2.2. Frequencies of words

The basic structure of a language consists of words. By analyzing the frequency distribution of words, one can draw valuable information about the functioning of communication system between people through language.



	word	%	sum
1	të	6.360	6.360
2	e	5.027	11.387
3	në	2.967	14.354
4	i	1.804	16.158
5	dhe	1.583	17.741
6	që	1.354	19.095
7	me	1.289	20.384
8	për	1.157	21.541
9	një	0.988	22.529
10	më	0.917	23.446
11	nuk	0.826	24.272
12	se	0.719	24.991
13	edhe	0.680	25.671
14	nga	0.677	26.348
15	si	0.645	26.993
16	është	0.590	27.583
17	do	0.546	28.129
18	ka	0.483	28.612
19	së	0.444	29.056
20	mund	0.374	29.430
21	u	0.354	29.784
22	po	0.322	30.106
23	por	0.300	30.406
24	ta	0.290	30.696
25	kur	0.283	30.979
26	këtë	0.280	31.259
27	shumë	0.257	31.516
28	ose	0.251	31.767
29	janë	0.245	32.012
30	tij	0.244	32.256
31	duhet	0.223	32.479
32	ai	0.221	32.700
33	kishte	0.206	32.906
34	duke	0.202	33.108
35	ishte	0.197	33.305
36	tyre	0.194	33.499
37	t'i	0.188	33.687
38	a	0.182	33.869
39	kanë	0.182	34.051
40	sa	0.177	34.228

Figure 4 – Distribution of the 40 most used words in texts in Albanian

After analyzing a large number of words (more than 38 million), written in different texts in Albanian, more frequent words of texts in Albanian are found. Thus, in diagram given in figure 4 are presented the results for the most frequent words of the Albanian language.

The diagram shows that most frequently used word (more than 6%) is the word "të", while after it, with frequency of about 5% , is placed the word (conjunction) "**e**", or both of them together constitute over 11% of words in Albanian texts. Frequency data, i.e. the order of words in this figure should not be understood as definitive. By analyzing other texts, or if is increase the number of texts used, these results are subject to change. However, it is important to note that the first 40 words (most used) constitute about 34% of all words written in Albanian. Whereas, with the 100 most frequently used words are written about 40% of all texts written in Albanian.

Frequency of most used words in Albanian texts has a special importance, because a very effective solution compiling applications for conversion of text into speech is to save these words as separate acoustic segments. As a result, quality generation of these words has impact on the quality of the speech synthesizing in Albanian, due to their greater "weight".

3. Designing of algorithm for synthesis

Given what was said above for Albanian language, during the speech synthesizing with the method of conjunction (concatenative synthesis) of acoustic segments stored in the database, firstly will be searched whether the requested word is found in the group of these words. If the word is found in this group, then the conversion will be "ideal". If the requested word is not in the group of words, then the conversion should be done by bringing together acoustic segments for twoletter. e characters. Otherwise, letters can be used as basis for generation, as it is shown in the figure 5.

```
repeat
{
   for each word
   {
      if word in base
            generate speach
      else
         for each twoletters
         {
            if twoletters in base
                  generate speach
            else
            {
            for each letter
                  generate speach
            }
         }
   }
}
until end of text
```

Figure 5 – Pseudocode for generation of speech for the given text in Albanian language

Before starting the generation process, the text is fixed: shortcuts are replaced with words, numbers are replaced with words, two-symbolic letters are replaced with special characters, and symbols such as $ or € are replaced with respective words, etc. Then, the system for the most frequent words and two-symbol letters is used, which initially generates the corresponding sound files that are stored in the database. Afterwards, begins the creation of sound file for the text, using the algorithm given in figure 5.

4. Conclusions

From what was said above, the options presented in various works, and from our work so far in testing possible solutions, we are creating a system which may give favorable results. For this purpose will be used algorithm in which are distinguished three main steps, as explained below.

The first step involves basic operations who adopt the text.

- *Initially the whole text is converted to fit so that two-symbols are replaced with single characters.*
- *Abbreviations that are included in the database are replaced with the texts as they are read/pronounced.*
- *Numbers are converted to text using words that are stored in the database.*

In the second step begins the process of replacement of textual segments with corresponding sound files, which will be placed in a final sound file. Here, is applied algorithm, which takes the specific words of the text and examines them under the following procedure.

- *If the word is found in the database, the appropriate sound file is taken and added to the final sound file.*
- *Otherwise, the word is fragmented into two-letter segments and for each segment is taken appropriate sound file and is placed in the final sound file.*
- *Similarly, is acted with numbers and abbreviations.*
- *If it happens that any digraph lacks in the database, sound files of separate characters/letters are used.*

We are analyzing the Albanian texts in order to find the larger units of text segments composed of two or more words, so that the relevant sound files are used, what will certainly affect the growth of naturality and clarity of the generated speech.

References

[1] Shaban Demiraj. **The Origin of the Albanians.** Academy of Sciences of Albania, Tirana, 2008
[2] Alan G. Konheim. **One-Gram Probability Distribution**. Cryptography – A Primer, John Wiley, 1981
[3] Agni Dika. **Distribucioni i frekuencave të shkronjave dhe grafemave në tekstet e shkruara shqipe.** Academy of Sciences and Arts of Kosova, Prishtina, 2006
[4] Bahri Beci, Ermira Topi. **Vëzhgime për dendurinë e përdorimit të fonemave zanore e bashkëtingëllore në gjuhën e sotme letrare shqipe.** Studime Filologjike 1, Tirana, 1986
[5] Mentor Hamiti, Agni Dika. **Learning opportunities through generating speech from written texts**. In *Procedia-Social and Behavioral Sciences Journal, Volume 2, Issue 2, pp. 4319-4324*. ELSEVIER, Istanbul, Turkey, 2010, ISBN 1877-0428.
[6] Daniel Jurafsky, James Martin. **Speech and Language Processing.** Pearson International Edition, 2008
[7] Agni Dika, Mentor Hamiti. **Options for Converting Albanian Written Texts into Speech.** *Proceedings of the Fourth Balkan Conference in Informatics, pp. 88-93.* A.T.E.I. Thessalonikis Publishing House, Thessaloniki, Greece, 2009. ISBN 978-960-287-127-0.

Agni Dika, PhD degree in Computer Sciences (1989) at the University of Zagreb, Croatia; Full Professor at the University of Prishtina, Faculty of Electrical and Computer Engineering.

Adnan Maxhuni, Master of Computer Science (2005) at the University of Prishtina; Teaching and Research Assistant at the University of Prishtina.

Avni Rexhepi, Master of Computer Science (2004) at the University of Prishtina; Teaching and Research Assistant at the University of Prishtina.