Available online at www.sciencedirect.com

ScienceDirect

IFAC-PapersOnLine 49-29 (2016) 259–262

Alternative Creation of Text to Speech Technology for the Albanian Language

Blerand Koshi* Xhevahir Bajrami**
Mentor Hamiti***

**University of Prizren "Ukshin Hoti", Prizren, Kosovo*
(Tel: +37744107945; e-mail: blerandkoshi@gmail.com).
***Vienna University of Technology, Institute for Handling Devices and Robotics (IHRT),*
Favoritenstrae 9/E325 A6, Vienna, Austria, (e-mail: xhevahirbajrami@gmail.com).
**** University of Prizren "Ukshin Hoti", Prizren, Kosovo,*
(e-mail: mentor.hamiti@uni-prizren.com)

Abstract: Nowadays there are ranges of technologies for converting texts into spoken language (speech). And their common aim is the artificial compilation of natural language - speech. Anyway, such quality converter still doesn't exist for any language much less for financially weak countries. As Text to Speech seems to be a great alternative for blind people, who can be informed about digital texts independently in any of their languages, the research in this field is more than necessary. The research is done by using Turkish language TTS for written texts in the Albanian language. The Turkish language is based on Latin letters and, the reading of them within sentences is similar to the reading in singularity. As the Albanian language has the same characteristics as well, the Albanian texts can use TTS of the Turkish language.

Keywords: Text, Speech, Generation, Albanian, Turkish, Language.

1. INTRODUCTION

Text-to-speech means the conversion of texts into clearly understandable speech. This helps the transferring of information from machine to a human being. (Bickley et al. 1998). The speech construction based on written texts, which is known as TTS (text-to-speech) although, has been improved continuously, yet it hasn't been solved. Various technologies which offer speech synthesis are used, such as: concatenate synthesis, formant synthesis, and articulatory synthesis. Certainly, these three types of technologies have their pros and cons. This kind of synthesis requires high-quality naturalness and intelligibility. So in all systems, the aim is to reach the perfection of both characteristics. Moreover, the technologies are affected by concrete languages where the written text is required to be synthesized. In this way, taking into consideration the differences in language, the speech synthesis for any new language is a peculiarity, and a universal application has not been found. Especially languages that belong to financially weak countries that have a small impact factor, have a great stagnation in this aspect, and unfortunately, the Albanian language is one of them.

For solving this issue, for the poor country languages we can use TTS for other languages, but of course by adopting it. The goal of this research is the measurement of opportunities, whereas as example was taken the text to speech for Turkish language for reading the texts in Albanian one, whereas the adjustment of the text will be done through normalization application made in C# programming language. The research is based on three main steps. The first step is finding the language that has similarities with the primary language, which in this case is the Albanian language. The second step is letter replacement based on the pronunciation. The final step is using/creating any application with TTS functions that can read installed speakers and can make the letter replacement in correct order.

2. THE CLASSIFICATION OF ALPHABET LETTERS

The classification of alphabet letters is specific for each language. In our case, the Turkish language has 29 letters, 22 of them Latin and 7 diacritic. The Albanian language has 36 letters, where beside Latin letters there are double letters and diacritic ones as well.

2.1 Albanian Language

As the majority of the world and European languages, the Albanian language as well had compiled its alphabet based in the Latin language. The Latin alphabet as we know consists of 26 letters such as: a, b, c, c, d, e, f, g, h, i, j, k, l, m, n, o, p, r, s, t, u, v, x, y, and z. All these letters are used in written English language, Italian and also in German and French, with some adjustments.

The combination of graphemes for the formation of special phonemes is not counted as real letters. For example, the English language is known as a refined language, where for interpreting 40 essential sounds are used 200 letter combinations. In contrast of these complex ways of writing, the Albanian language uses the Latin alphabet, but tailored to its phonemic inventory. For interpreting all phonemes of spoken Albanian language, a total of 36 letters are used. Since the Latin alphabet has only 26 letters, they aren't

10.1016/j.ifacol.2016.11.063

enough to write 36 letters of the Albanian language, that's why 25 letters (without ë) are filled with other letters. This issue was resolved in the Congress of Manastir in November 1908, where the letters of Albanian language were defined and ordered as follows (Instituti i Gjuhësisë dhe i Letërsisë, 1995): a, b, c, ç, d, dh, e, ë, f, g, gj, h, i, j, k, l, ll, m, n, nj, o, p, q, r, rr, s, sh, t, th, u, v, x, xh, y, z, zh.

Thus, if we evaluate the graphic system of Albanian language according to general principles, we can freely say that is one of the easiest (comparing with others alphabets). Moreover, the problem of writing was solved very properly: 36 phonemes of Albanian are interpreted with 36 special letters, which are made by 25 simple letters, 9 double letters and two letters containing diacritic signs (ë and ç). In this aspect we can say that the basic graphic principle is respected, as the same phoneme is introduced in all possible cases, using only one special letter (simple or double one). Instituti i Gjuhësisë dhe i Letërsisë, (2002).

2.2 Turkish Language

Turkish alphabet is constructed based on Latin letters, it consists of 29 letters, where letters as 'b', 'c', 'ç', 'd', 'f', 'g', 'ğ', 'h', 'j', 'k', 'l', 'm', 'n', 'p', 'r', 's', 'ş', 't', 'v', 'y', 'z' are consonants whereas 'a', 'e', 'ı', 'i', 'o', 'ö', 'u' and 'ü' are vowels. Also, the Turkish language in some cases uses (^) above letters as: a, e, ı, i, o, ö, u and ü vowel, to explain the length while reading these letters. This alphabet represents modern Turkish pronunciation with a high degree of accuracy and specificity. It is the current official alphabet and the latest in a series of distinct alphabets used in different eras. The current 29-letter Turkish alphabet was established as a personal initiative of the founder of the Turkish Republic, Mustafa Kemal Atatürk.

2.3 Main Distinctions Between Two Languages

Probably the Albanian and the Turkish languages have many differences, but since the paper is about TTS, only the letters, and their pronunciation are considered. Even though both languages are based on the Latin language and are very similar to each other in reading, not all letters fulfil this criterion. Letters which almost are read the same in the two languages are: a, b, ç, d, e, f, g, h, i, k, l, m, n, o, p, r, s, t, u, v, z. Since the Turkish alphabet doesn't have double letters, there will be a difficulty in adapting. The main distinction that can be mentioned here is the number of letters, where the Albanian language has 7 letters more than the Turkish one. Letters like ğ and ö don't resemble with any of Albanian letters. On the other side letters like, dh, the, ll and x don't resemble with any of the Turkish language letters.

3. THE NORMALIZATION OF TEXT

The normalization of text, in general, represents the process of transforming the text in any certain form, according to previously defined rules, with the aim of further processing the text with ease. Usually, the normalization precedes the process for certain destinations. Since such cases may include: speech synthesis, automated language translation, the storage of data based on the respective normalized texts, etc.

As examples of normalization, the following actions can be mentioned:

- Normalization based on UNICODE,

-The conversion of all letters within text in small letters and capital ones,

-The omission of punctuation marks within text,

- The replacement of abbreviations with full proper names, etc.

Usually, while choosing the actions which are used for the normalization of texts, except adjustments we also should take in consider the language in which the text was written. For different languages there are also compiled different normalization procedures.

This is conditioned by the specifics of the respective languages. For example, there are languages that use the Cyrillic alphabet, some languages do not read text letter for letter as it is written, some languages use characters with diacritical marks, etc. So, the normalization of texts written in the Albanian language will be made according to use its features in Turkish, trying maximally to achieve the best possible results.

3.1 The Normalization of Written Text in the Albanian Language

The differences between the languages normalization of text written in the Albanian language will be made by adapting several related letters with the Turkish language. As some of the letters will be read correctly, the rest will remain to be the best solution possible. During normalization special attention is paid to priority letters. Some character pronunciation such as 'c' in Turkish is close to 'GJ' and 'XH' in Albanian, and before replacing them with 'c', the first substitution that was made is 'c' with 's'. Other substitutions are shown in Table 1, where the first priority marked letters or rather the letters that should be normalized before others. Special characters such as $, €, +, = are translated at the beginning in the Albanian language letters and then are adapted to Turkish. Also, the numbers 0-9 normalization is made in the same way, where the numbers will normalize in a digital basis (100 = one double zero).

Table 1. Replacement of Albanian letters with Turkish ones

Letter	Replacement	Letter	Replacement
c	s	y	ü
xh	c	j	y
gj	c	zh	j
x	s	rr	r
dh	v	sh	Ş
ë	ı	th	f
ll	l	q	ç