SANDRA SOVILJ-NIKIC
MILAN SECUJSKI
VLADO DELIC
IVAN SOVILJ-NIKIC
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
sandrasn@eunet.yu

Analysis of different factors influencing vowel duration in the Serbian language

1. INTRODUCTION

Being a very perspective speech technology, speech synthesis has been thoroughly studied over the past few years at the Faculty of Technical Sciences in Novi Sad in Serbia as part of the AlfaNum project. One of the main goals of this project is the design of a high-quality speech synthesizer for the Serbian language. A high-quality speech synthesizer is one with an ability to produce intelligible synthesized speech which sounds natural. One of the basic requirements of the TTS (text-to speech) system is for synthesized speech to be natural which also influences the intelligibility of synthesized speech. In the many existing TTS systems the naturalness of synthesized speech is often very poor, because achieving it is a very difficult task. As a result current research that is part of the AlfaNum project is directed at the improvement of naturalness of synthesized speech produced by the existing speech synthesizer for the Serbian language developed at the Faculty of Technical Sciences.

The results presented in this paper obtained on account of statistical analysis of the different factors influencing vowel duration in the Serbian language will contribute to the improvement of naturalness of synthesized speech. In this paper the AlfaNum synthesizer and the speech database used for analyzing vowel duration are also described. The results of the analysis related to the duration of a vowel depending on its neighboring phonemes and their voicing are presented. Also, the results of the analysis showing the influence of the position of a stressed vowel within a word on the duration of unstressed vowels are discussed.

2. THE ALFANUM SPEECH SYNTHESIZER

The task of any TTS system is to produce a speech signal which will be intelligible and will sound natural to a human listener. The text-to-speech synthesis procedure consists of two main phases. The first one is text analysis, where the input text is transcribed into phonetic or some other linguistic representation, and the second one is the generation of speech waveforms where the acoustic output is produced from the phonetic and prosodic information. These two phases are usually called as high- and low-level synthesis (Figure 1).

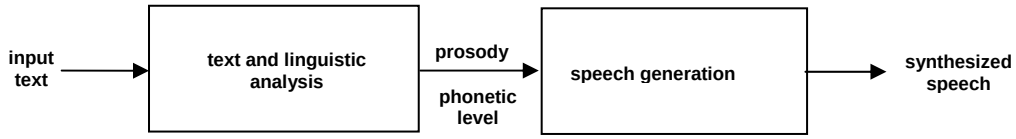

Figure 1. Text-to-speech synthesis procedure

The high-level synthesis can be divided in three main phases: text preprocessing, pronunciation analysis and prosodic analysis. Text preprocessing is the phase in which numerals, special characters; abbreviations and acronyms are expanded into full words. Pronunciation analysis is the determination of the pronunciation of certain words, i.e. obtaining the phonetic transcription. In the Serbian language this is quite simple because written text almost corresponds to its pronunciation. The main problem for the TTS system in the Serbian language is the text accentuation which it has to perform using the comprehensive accent dictionaries as well as methods for morpho-syntactic sentence analysis. These methods should be able to solve problems that emerge, such as when certain words in a sentence are accentuated in a several different ways. The prosodic features of speech are determined in the last phase of high-level synthesis called prosodic analysis. Prosodic features can be divided into several levels such as a syllable, word or phrase level and they consist of pitch, duration and stress over the time. The prosody of continuous speech depends on many different features, such as the meaning of the sentence, the characteristics of the speaker and emotions (Figure 2). Unfortunately, written text does not contain information of prosodic features, some of which change dynamically during speech. Thus, the modeling of prosodic features is a very complex problem. Solving of this problem is the subject of a lot of research and experiments especially having in mind the great impact of prosody on the naturalness of synthesized speech.

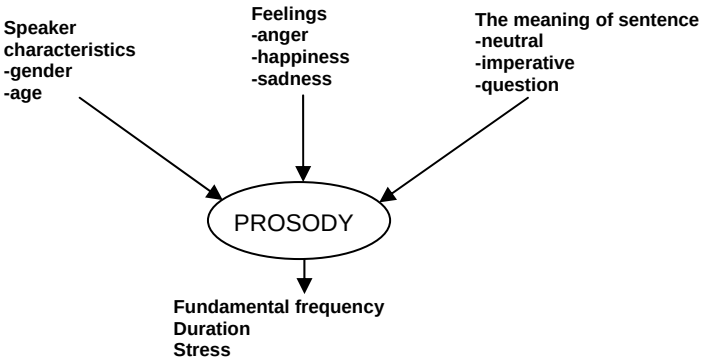

Figure 2. Prosodic dependencies

Synthesized speech can be produced using several different methods following text and prosodic analysis [1]. All of these methods have some benefits and problems of their own. The methods are usually classified into three groups:

- articulatory synthesis,
- formant synthesis,
- concatenative synthesis.

Concatenative synthesis uses different lengths of prerecorded samples derived from natural speech. This is probably the easiest way to produce intelligible and natural sounding synthetic speech. This method is the most commonly used one in many TTS systems. However, concatenative synthesis is usually limited to one speaker and requires more memory capacity than other methods. Our AlfaNum synthesizer is a concatenation text-to-speech synthesizer based on TD-PSOLA (Time Domain Pitch-Synchronous OverLap-and-Add) algorithm carried out on speech segments which are on-line selected from prerecorded speech database. On-line selection of speech segments were grouped according predefined criteria aimed at minimizing audible discontinuities leads to fairly intelligible and natural-sounding synthetic speech. Off-line preprocessing of the speech corpus, such as labeling the database and its pitch-marking, is necessary for implementation of TD-PSOLA algorithm.

Time domain version, TD-PSOLA, is the most commonly used due to its computational efficiency. The basic algorithm consists of three steps. The analysis step where the original speech signal is first divided into separate but often overlapping short-term analysis signals, the modification of each analysis to a synthesis signal, and the synthesis step where these segments are recombined by means of overlap-adding.

3. THE SPEECH DATABASE

3.1 Contents

The speech database contains approximately two hours of continuous speech pronounced by a single female speaker. Having in mind the possible applications of such a system, and the fact that concatenation of longer speech segments yields more intelligible and more natural speech, it was decided that the database should include phrases such as commonplace first and last names, addresses, names of companies, cities and countries, amounts of money, currencies, time and date phrases, horoscope reports, weather reports, typical phrases used in interactive voice-response systems, in e-mail messages etc. The database was recorded in laboratory conditions, and submitted to several off-line operations necessary for implementing TTS, such as labeling the database and its pitch-marking.

3.2 Labeling and Pitch-Marking

The labeling of the database consists of placing boundaries between units belonging to a previously established set of units such as phonemes. It actually entails storing of information about units, such as starting and ending moment, in a separate database. The labeling of the AlfaNum speech database was predominantly phoneme based, although in some cases a better alternative was adopted, due to certain phonetic features of particular phones, as well as certain peculiarities of Serbian language. For instance, some classes of phones, such as plosives and affricates, were considered as pairs of semiphones (including occlusion and explosion in case of plosives and occlusion and friction in case of affricates). Also, depending on the stress type, the timbre of the five vowels of Serbian language can vary. Due to difficulties in modifying vowel timbre without use of more computationally intensive parametric speech synthesis algorithms, and as a result avoiding timbre mismatches that lower quality of synthesized speech, the solution we adopted was defining classes of distinctive timbre variants of each vowel, and considering them as different vowels altogether. This is one of the criteria for on-line selection of segments. Considering the way the phoneme R is pronounced in Serbian language (a periodical set

of occlusions and explosions produced by tongue oscillating against the hard palate), all of these occlusions and explosions were treated as distinct phonemes.

Labeling was performed automatically, using the AlfaNum continuous speech recognizer [6], and verified by a human expert. Verification was based on signal waveform, its spectrogram and its auditory perception. SpeechVis software previously developed within the AlfaNum project was used [7].

Implementing TD-PSOLA algorithm involves prior pitch-marking of the database, that is detecting locations within phones most suitable for centering overlapping windows and extracting frames, in case a TTS algorithm which requires pitch-synchronous frame positioning should be used. Thus, during voiced frames, one marker per period was appointed, and during unvoiced frames, markers were appointed according to the average fundamental frequency throughout the database. In order to avoid audible effects caused by abrupt changes in V/UV marker positioning strategies, UV positioning strategy was somewhat modified in the vicinity of V/UV boundaries, and thus the rate with which distances between adjacent markers can vary was severely reduced.

4. EXPERIMENTAL RESULTS

As aforementioned automatic derivation of prosodic information from plain text is one of the fundamental problems of speech synthesis. The task of the prosody generator is to perform automatic stress assignment at word and sentence level and to produce a set of prosodic parameters (f_0 curve, phoneme durations and energy curve) that would yield natural-sounding synthesized speech.

Serbian is an Indo-European, South-Slavic language with 10 million speakers in Serbia and Montenegro (11 million world-wide). There are several important aspects in speech synthesis where some peculiarities of the Serbian language should be taken into account. A complex accentuation system in Serbian is of special interest for accurate prosody generation. The five vowels in the Serbian language can each be stressed in four different ways, according to the pitch level during the stressed vowel itself, its relation to the pitch level of the next syllable, and the duration of the stressed syllable, which falls into two classes: long and short. Four different types of stress can be recognized as rise/long, rise/short, fall/long and fall/short. The pitch level carried by a word is an essential feature of the meaning of that word. Also, vowel timbre can vary with accent type. Thus, vowels belonging to classes with significantly different timbres are considered as different vowels altogether. According to the rule, vowels stressed with a short accent type have more opened timbre and vowels stressed with a long accent type have more closed timbre.

In a current version of the AlfaNum speech synthesizer, f_0 curve was treated as a prosodic parameter having the most important influence on word and sentence prosody in the Serbian language. In accordance with that, the other prosodic parameters and their impact on prosody were not studied in detail within our project until now. The prosody generator used consonant durations calculated for three distinctive positions (initial, medial and final) for every consonant and a number of the most frequent consonant groups, as well as vowel durations calculated for the same three positions, for unstressed and stressed vowels.

Within this paper in order to improve the naturalness of synthesized speech more detailed statistical analysis of vowel duration in a large speech database was performed. Software for analyzing the speech database was written in Visual Studio C++ programming language. The influence of neighboring phonemes and their voicing on vowel duration was analyzed. The results of this analysis show that vowels in a voiced neighborhood have a longer duration than vowels which neighboring phonemes are unvoiced. Also, obtained results show that different vowels have significantly different durations. In order to improve the naturalness of synthesized speech, this fact should be taken into account too. The results are given in Tables 1, 2, 3, 4, 5. The values refer to the duration of vowels in general, stressed vowels, stressed vowels with opened timbre, stressed vowels with closed timbre and unstressed vowels, in voiced and unvoiced neighborhood separately. Similar results were obtained for the Slovenian language in the Laboratory of Artificial Perception at the University of Ljubljana [3]. Slovenian is also South-Slavic language as the Serbian.

VOWEL	DURATION OF VOWEL [ms] IN VOICED NEIGHBORHOOD	VOWEL	DURATION OF VOWEL [ms] IN UNVOICED NEIGHBORHOOD
A	119.21	A	113.31
E	108.2	E	96.26
I	98.55	I	78.64
O	109	O	104.33
U	112.96	U	103.12
AVERAGE	109.58	AVERAGE	99.13

Table 1. Duration of vowels in general

VOWEL	DURATION OF VOWEL [ms] IN VOICED NEIGHBORHOOD	VOWEL	DURATION OF VOWEL [ms] IN UNVOICED NEIGHBORHOOD
A	170.6	A	144.92
E	146.58	E	143.9
I	127.84	I	109.07
O	149.39	O	129.33
U	142.49	U	122.23
AVERAGE	147.38	AVERAGE	129.89

Table 2. Duration of stressed vowels

VOWEL	DURATION OF VOWEL [ms] IN VOICED NEIGHBORHOOD	VOWEL	DURATION OF VOWEL [ms] IN UNVOICED NEIGHBORHOOD
A	153.91	A	133.3
E	136.88	E	134.07
I	111.79	I	101.1
O	143.19	O	129.38
U	117.28	U	109.04
AVERAGE	132.47	AVERAGE	121.37

Table3. Duration of stressed vowels with opened timbre

VOWEL	DURATION OF VOWEL [ms] IN VOICED NEIGHBORHOOD	VOWEL	DURATION OF VOWEL [ms] IN UNVOICED NEIGHBORHOOD
A	185.34	A	186.06
E	168.11	E	159.21
I	140.53	I	127.23
O	172.09	O	147.32
U	175.75	U	139.75
AVERAGE	168.36	AVERAGE	151.91

Table 4. Duration of stressed vowels with closed timbre

VOWEL	DURATION OF VOWEL [ms] IN VOICED NEIGHBORHOOD	VOWEL	DURATION OF VOWEL [ms] IN UNVOICED NEIGHBORHOOD
A	80.05	A	70.44
E	75.74	E	57.3
I	77.3	I	55.98
O	75.34	O	65.41
U	77.92	U	63.88
AVERAGE	77.27	AVERAGE	62.6

Table 5. Duration of unstressed vowels

The remaining part of the analysis performed within this paper show that vowel position relative to the stressed vowel within the word has an impact on the duration of unstressed vowels. The results are presented in Figure 3. The mean represents a normalized average value of the ratio of durations of unstressed vowels which are at different positions referring to the stressed vowel within the word. Annotated with I is the ratio of unstressed vowels which are on the third and the first position before the stressed syllable, annotated with II is the ratio of the second and the first vowel before the stressed syllable, III is assigned to the ratio of the first vowel before the stressed syllable and the first vowel after the stressed syllable, IV is assigned the ratio of the first and second vowel after the stressed syllable and finally V is assigned to the ratio of the first and the third vowel after the stressed syllable. Normalization with an average duration of the given vowel in the given neighborhood was performed. The average durations were obtained as a result of the first part of the analysis within this paper. On the basis of the given results it can be concluded that the closer the unstressed vowels to the stressed vowel the longer their duration.

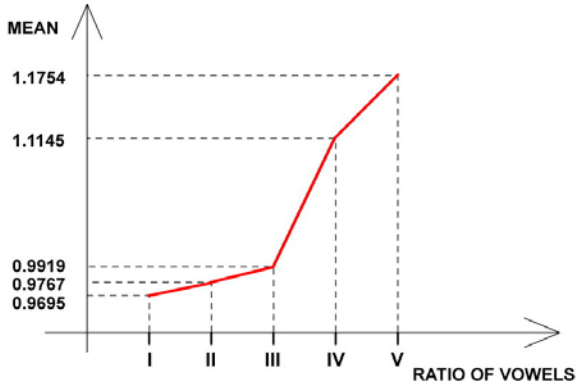

Figure 3. The influence of a position of stressed vowel on the duration of unstressed vowels

5. CONCLUSION

In this paper the concept of the AlfaNum concatenation-based speech synthesizer in the Serbian language as well as large speech database containing continuous speech are described. In order to improve the naturalness of synthesized speech produced by the existing speech synthesizer a statistical analysis of vowel duration in a large speech database was performed. The implementation of results presented in this paper will contribute to the improvement of naturalness of synthesized speech produced by the AlfaNum synthesizer which is the first complete and commercially applicable TTS in Serbian. There are many applications in which highly intelligible and natural-sounding synthesized speech can be used such as speech-enabled telecommunications services (voice portals, interactive voice-response systems, automatization of call centers, SMS and e-mail reader), language education and aids for persons with disabilities.

REFERENCES

[1] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers. Dordrecht/ Boston/ London, 1997.

[2] M. Secujski, R. Obradovic, D.Pekar, Lj. Jovanov, V. Delic, “AlfaNum System for Speech Synthesis in Serbian Language”, *in Proc. 5-th Conf. Text, Speech and Dialogue*, Brno, Czech Republic, 2002, pp. 8-16.

[3] J. Gros, *Samodejno tvorjenje govora iz besedil*, Zalozba ZRC, Ljubljana, Slovenia, 2000.

[4] S. Jovicic, *Govorna komunikacija, fiziologija, psihoakustika i percepcija,* Nauka, Belgrade, Serbia, 1999.

[5] S. Lemmetty, “Review of Speech Synthesis”, Master thesis, University of Technology, Helsinki, Finland, 1999.

[6] D. Pekar, R. Obradovic, V. Delic, “AlfaNumCASR - A Continuous Speech Recognition System”, DOGS, Becej, Serbia, 2002.

[7] R. Obradovic, D. Pekar, “C++ Library for Signal Processing – SLIB”, DOGS, Novi Sad, Serbia, 2000.