AlfaNum System for Speech Synthesis in Serbian Language

Milan Sečujski, Radovan Obradović, Darko Pekar,
Ljubomir Jovanov, and Vlado Delić

Faculty of Engineering, University of Novi Sad, Yugoslavia
{secujski, tlk_delic}@uns.ns.ac.yu

Abstract. This paper presents some basic criteria for conception of a concatenative text-to-speech synthesizer in Serbian language. The paper describes the prosody generator which was used and reflects upon several peculiarities of Serbian language which led to its adoption. Within the paper, the results of an experiment showing the influence of natural-sounding prosody on human speech recognition are discussed. The paper also describes criteria for on-line selection of appropriate segments from a large speech corpus, as well as criteria for off-line preparations of the speech database for synthesis.

1 Introduction

Being a very prospective speech technology, speech synthesis has been thoroughly studied at the University of Novi Sad, Yugoslavia, within the AlfaNum project, for several years. One of the main goals of this project is the design of a high-quality concatenation-based speech synthesizer for Serbian language. The first version of the AlfaNum synthesizer is based on TD-PSOLA algorithm, performed on segments selected on-line from a large speech database containing continuous speech, according to various criteria defined beforehand. This paper gives a detailed description of the AlfaNum synthesizer, including descriptions of a dictionary-based prosody generation module, and implementation of on-line selection of segments. Off-line preprocessing of the speech corpus which was necessary for implementation of TD-PSOLA algorithm is also described.

2 On Some Peculiarities of Serbian Language

In this paper we will discuss some peculiarities of Serbian language that are relevant for speech synthesis. There are several important aspects in speech synthesis where those peculiarities should be taken into account.

One of the most remarkable features of Serbian language is its most simple grapheme-to-phoneme conversion rule. Due to a radical language reform carried out in the 19th century, each letter today corresponds to exactly one sound. The exceptions to this rule are exceedingly rare and most of them can occur only at the boundaries between two words, where a voiced consonant can turn into its

voiceless sibling if followed by a voiceless consonant and vice-versa, which makes detection of such exceptions trivial. Several phonemes have their allophones, but the only one that significantly affects speech intelligibility is the "velar n", a sound very similar to English n appearing in word such as 'song'. This allophone features only before velar stops, and is therefore quite simple to predict. Thus, the task of phonetization in Serbian speech synthesis is reduced to a trivial check, and solutions based on dictionaries and morphophonemic rules are not needed.

The task of a prosody generator is to ensure that the pronounced phrase sounds as naturally as possible. Generally, beside being more agreeable to a human listener, natural-sounding synthesized speech is easier to understand inasmuch as it is easier to perform lexical segmentation upon it, that is, to identify boundaries between words. In Serbian language, the importance of natural prosody is even more emphasized, since the location of stress within words is sometimes an essential feature of the meaning of that word. In order to confirm that natural prosody is of great importance for lexical segmentation and therefore for understanding, especially in adverse conditions, an experiment was conducted, which will be described later.

The five vowels of Serbian language can be stressed in four different ways each, according to pitch level during the stressed vowel itself, its relation to the pitch level of the next syllable, and duration of stressed syllable, which falls into two classes: long and short. Four different types of stress can thus be recognized as rise/long, rise/short, fall/long and fall/short, as shown in Figure 1. Depending on stress types, timbre of these vowels, as well as formant structure, can also vary. Due to difficulties in modifying vowel timbre without use of more computationally intensive parametric speech synthesis algorithms, and thus avoiding timbre mismatches that lower the 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 which must be kept in mind.

Another remarkable feature of Serbian language is characteristic of most other Slavic languages as well. Namely, the consonant R can serve as a vowel in case it is located between two other consonants. When pronounced as such, it is preceded by a vowel sound similar to one appearing in English word "burn". That sound is omitted in written Serbian language, but it is quite easy to reconstruct its position. All prosody modifications performed upon R serving as a vowel are actually performed on this vowel sound. All these facts must be taken into account not only when performing grapheme-to-phoneme conversion, but also when labeling the speech database.

3 The Speech Database

3.1 The Contents

The speech database contains approximately two hours of continuous speech, pronounced by a single female speaker. Having in mind possible applications of

Fig. 1. Stress types in Serbian language and corresponding f_0 contours: (a) rise/long, (b) rise/short, (c) fall/long, (d) fall/short, as pronounced by the speaker involved in database recording

such a system, and the fact that concatenation of longer speech segments yields more intelligible 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, weather reports, horoscope reports, typical phrases used in interactive voice-response systems, typical phrases used 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 implies storing information about units 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). Vowels belonging to classes with significantly different timbres were considered as different vowels altogether, and therefore not

interchangeable. Considering the way the phoneme R is pronounced in Serbian language (a periodical set of occlusions and explosions produced by the 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 [7], and verified by a human expert. Verifying was based on signal waveform, its spectrogram and its auditory perception. The SpeechVis software, previously developed within the AlfaNum project, was used [8].

Implementing TD-PSOLA algorithm implies previous 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.

As to positioning pitch markers within voiced frames, the process was carried out in two phases. To begin with, preliminary estimations of pitch contours of each segment were made using AMDF pitch-extraction method. Each of the segments was previously low pass filtered with cutoff frequency 900 Hz. The next step was locating the frame with the highest degree of voicing and locating the maximum peak within that frame. The initial pitch marker was placed there. Afterwards, the search for other pitch markers was conducted according to preliminary pitch estimations, which resulted in placing pitch markers in such a way that windows centered around them would cover most of the waveform's energy, and no significant distortion of the signal caused by windowing could occur.

Such a procedure is not entirely error-free, because low pass filtering can sometimes modify peak values to such an extent that peaks recognized as maximum in some segments do not coincide with peaks recognized as maximum in the rest. This can happen in case of voiced phones whose waveforms have two prominent peaks of roughly the same height. There is another reason why errors of this kind may occur. At the precise spot of boundaries between phones there is sometimes an irregularity in functioning of the glottis, leading to a discontinuity in positions of glotal impulses. If a phoneme-based pitch-marking is adopted, that is, if pitch-marking is performed independently within phones, and not within speech segments containing more than one phone, most of such errors are eliminated completely.

The database containing information about pitch markers is suitable for speech synthesis using any of the concatenation-based techniques, even in case of techniques that do not require explicit knowledge of pitch markers, since pitch contours can be determined from pitch marker positions in a straightforward way.

4 Prosody Generation

Acoustic parameters such as f0 contour were calculated in two steps. The first step consisted of analysing the sentence, word by word, in order to get information such as stress types and locations, part of speech classes and functions of particular words in the sentence. Grammatical information is essential for synthesis of natural-sounding prosody, since words in Serbian language can sometimes be stressed in different ways depending on their morphologic categories, and sometimes even have different meanings if stressed differently. In some cases even syntactic analysis does not help, and several interpretations of the same sentence, all of them grammatical, can be stressed in different ways and therefore yield different meanings. For a human listener there is no confusion, because he/she relies on contextual information.

4.1 The Dictionary

Since stress in Serbian language is fairly unpredictable, a dictionary-based solution was adopted. A special dictionary including information on stress configuration, part of speech class and morphologic categories for each word was created. Furthermore, since stress can vary along with inflections of the same word, and those variations are predictable only to a certain extent, it was necessary to include all word forms as separate entries in the dictionary. Several part of speech classes were identified as having regular behaviour when submitted to inflection and they were entered in the dictionary in a form which occupied little space, but was sufficient for correct determination of the stress of every inflected form. In such a way the dictionary containing more than one million entries (including inflected forms) occupied about 6.5 MB in txt format.

Such a solution is not entirely error-free, since it does not include syntactic analysis, nor does it solve cases when syntactic ambiguities arise, and semantic analysis, however primitive, must be performed. It was decided to leave these two problems for later stages of the project. Syntactic analysis, when implemented, would rely on information from the dictionary, and semantic analysis will be limited to checking up collocations in the dictionary – that is, deciding in favour of words that typically occur in particular contexts related to other words, rather than in favour of words that do not. The information on collocations in Serbian language will be acquired through statistical analysis of very large textual databases, and entered into the dictionary along with other information.

Another problem that occurs is that some words may not be found in the dictionary. It can happen because of their rarity, because a nonstandard affix was used, but also frequently in case of foreign names, names of companies etc. In that case, strategies for determining the correct way of stressing must be defined. Strategies currently being developed within our projects include making analogies based on standard prefixes and suffixes and rhyming.

The graphical user-interface created for entering words in the dictionary is dialog-based and highly intuitive. The person entering the dictionary must be familiar with lexis of stress system in Serbian, and must be able to stress words

properly. However, after a short introduction, even a lay user of the TTS system is able to add words to the dictionary if desired. For instance, (s)he might want to enter names of employees in his/her call centre.

4.2 F0 codebook

Using information from the dictionary, the system is able to reconstruct a particular stress configuration of a group of words which form a metrical unit. In this phase of the project, several f0 contours are assigned to each metrical unit, depending on its position in the sentence (beginning, neutral, before comma, ending), and the resulting f0 contour is smoothed in order to avoid audible pitch discontinuities and tilted towards the end of the sentence [1], [3]. The curves were extracted from typical stress contexts. Such a method does not take into account syntactical information, but relies only on punctuation marks. However, results are still significantly better than in case of synthesizers with constant f0, available in Serbian until now.

4.3 An Experimental Confirmation of the Influence of Prosody on Human Speech Recognition

Within this paper an experiment was conducted in order to show the influence of natural prosody on human speech recognition, that is, to show that the listener relies on prosody to a considerable extent, when required to reconstruct the syntactically and semantically correct sentence that (s)he has heard.

To that purpose, the AlfaNum synthesizer was used to create 12 syntactically and semantically correct sentences, whose length did not exceed 8 words. Those sentences were intended for recognition by human listeners. Each of those sentences was created in three variants. The first variant had completely flat f0 curve and equal durations of all vowels, which yielded metallic-sounding unnatural speech. The second variant had incorrect prosody features, based on wrong sentence accentuation, and the third had the correct prosody features.

Considering that the intelligibility of all three variants was relatively high in normal conditions, the experiment was conducted in adverse conditions. Gaussian and impulse noise were introduced into synthesized sentences, and the experiment took place in the presence of intensive ambiental noise.

The experiment included 12 listeners, and each of them listened to 12 sentences – four sentences from the first group, four from the second group and four from the third. Each of the listeners was required to repeat the sequence that (s)he has heard, and in case of incorrect recognition, they were required to repeat the recognition after hearing the sentence again. For each listener and each sentence it was recorded whether the sentence was correctly recognized at once, or after being repeated, or not recognized at all. Each of the listeners was previously informed that the sentences that they were about to hear were synthesized, and to each of them Serbian was the mother tongue.

The results of the experiment are shown on Figure 3, and they clearly show the importance of natural prosody. Another important result of the experiment

is that the sentences with wrong prosody were harder to recognize than sentences with flat f0 curve. This result also shows how much the listeners rely on prosody. When the f0 curve is flat and all the vowels have the same durations, the listeners are aware that the prosody does not yield any information and concentrate their efforts on phoneme recognition and combining phonemes into meaningful words, and ultimately, into a meaningful sentence. When variations of f0 and vowel durations are present, the listeners are under a wrong impression that the sentence is pronounced with a proper f0 curve and try to spot meaningful words by the way they sound, but fail in most cases. Therefore, results are worse than in case there were no f0 and vowel duration variations at all.

This experiment yields another conclusion. Considering that f0 curves used in this experiment were not recorded from actual sentences pronounced by a human speaker, but were combined from f0 curves taken from the codebook, the experiment can also be interpreted as a confirmation that the codebook is adequate.

5 On-line Selection of Segments

Halfphones are considered as basic units which cannot be further segmented, but it is desirable to extract segments as large as possible, in order to preserve intelligibility. According to differences between existing and required values of parameters previously defined, each speech segment which can be extracted and used for synthesis is assigned target cost, and according to differences at the boundaries between two segments, each pair of segments which can be concatenated is assigned concatenation cost [2]. Target cost is the measure of dissimilarity between existing and required prosodic features of segments, including duration, f0, energy and spectral mismatch. Concatenation cost is the measure of mismatch of the same features across unit boundaries. Various phoneme groups are treated in different fashion. Some types of phonemes, such as unvoiced plosives, are more suitable for segmentation than the others, and have lower concatenation costs. The degree of impairment of phones is also taken into account.

The task of the synthesizer is to find a best path through a trellis which represents the sentence, that is, the path along which the least overall cost is accumulated. The chosen path determines which segments are to be used for concatenation, as shown on Figure 2.

6 Conclusion

In this paper conception of high-quality TTS in Serbian language is described in detail. Prosody generation principles are presented, in view of several distinctive features of Serbian language. On-line selection of speech segments to be concatenated according to predefined criteria aimed at minimizing audible discontinuities leads to fairly intelligible and natural-sounding synthetic speech. After several previous attempts at creating a TTS system in Serbian which were

mostly diphone-based and did not treat prosody in any way, this is the first complete TTS in Serbian which is commercially applicable.

References

1. T. Dutoit: "An Introduction to Text-to-Speech Synthesis". Kluwer Academic Publishers, Dordrecht/Boston/London (1997)
2. M. Beutnagel, M. Mohri, M. Riley: "Rapid Unit Selection from a Large Speech Corpus for Concatenative Speech Synthesis". Proceedings of EUROSPEECH '99, pp.607-610. Budapest, Hungary (1999)
3. I. Lehiste, P. Ivić: "Word and Sentence Prosody in Serbocroatian". The Massachusetts Institute of Technology (1986)
4. Slobodan T. Jovičić: "Speech Communication, Physiology, Psychoacoustics and Perception". Izdavačko preduzeće NAUKA, Beograd, Yugoslavia (1999)
5. V. Delić, S. Krčo, D. Glavatović: "Basic Elements for ASR and TTS in Serbian Language". DOGS, pp. 32-37, Fruška Gora, Yugoslavia (1998)
6. M. Sečujski: "Text-to-Speech with Respect to Serbian Language". Graduation thesis, School of Engineering, Novi Sad, Yugoslavia (1999)
7. D. Pekar, R. Obradović, V. Delić: "AlfaNumCASR - A Continuous Speech Recognition System". DOGS, Bečej, Yugoslavia (2002)
8. R. Obradović, D. Pekar: "C++ Library for Signal Processing – SLIB". DOGS, Novi Sad, Yugoslavia (2000)