Automatic Prosody Generation for Serbo-Croatian Speech Synthesis Based on Regression Trees

Milan Sečujski[1], *Darko Pekar*[2], *Nikša Jakovljević*[1]

[1] Faculty of Technical Sciences, University of Novi Sad, Serbia
[2] "AlfaNum – Speech Technologies Ltd.", Novi Sad, Serbia
{secujski,jakovljevic}@uns.ac.rs, darko.pekar@alfanum.co.rs

10.21437/Interspeech.2011-790

Abstract

The paper presents the module for automatic generation of prosodic features of synthesized speech, namely, f_0 targets and phonetic segment durations, within the speech synthesizer AlfaNumTTS, the most sophisticated speech synthesis system for Serbo-Croatian language to date. The module is based on regression trees trained on a studio recorded single speaker database of Serbo-Croatian. The database has been annotated for phonemic identity as well as a number of prosodic events such as pitch accents, phrase breaks and prosodic prominence. Besides the traditional description of the intonational phonology of Serbo-Croatian through four distinct accent types, within this study we have examined the possibility of representing them as tonal sequences, which has been suggested in recent linguistic literature. The results obtained confirm that the four accents can indeed be reduced to sequences of high and low tones without loss of quality, provided that phonemic length contrast is preserved.

Index Terms: speech synthesis, Serbo-Croatian language, prosody generation, regression trees

1. Introduction

AlfaNumTTS is a concatenative speech synthesizer for Serbo-Croatian language developed at the Faculty of Technical Sciences, University of Novi Sad, Serbia[1] [1]. The paper will present the module for automatic prediction of prosodic features of synthesized speech based on regression trees, used within the AlfaNum speech synthesizer. The concept of classification and regression trees (CART) [2] has so far been successfully used to treat this problem for a variety of languages (cf. e.g. [3] and [4] for f_0 prediction and [5-9] for phonetic segment duration prediction). Being a less-resourced language, and one of the few pitch-accent languages of Europe, Serbo-Croatian has proven to be especially challenging from the point of view of the development of spoken language technologies.

The remainder of the paper is organized as follows. Section 2 will describe the Serbo-Croatian language with emphasis on its particularities most relevant to automatic prosody prediction. Section 3 will give a brief description of AlfaNumTTS as well as speech and language resources it relies on. The speech database used for the experiment as well as the conventions used in its annotation will be described in detail. Section 4 will present the system for automatic prosody prediction for Serbo-Croatian based on regression trees, including a comparison of the traditional paradigm of four distinct lexical pitch accents and two novel approaches suggesting that they can be effectively represented as tonal sequences. This section will also include the results of subjective and objective evaluation of the quality of synthesized speech carrying automatically generated prosody. The concluding section will discuss the results and give an outline of the future work.

[1] Although the synthesizers for both Eastern and Western standard variants of the language have been developed within our research, the speech database and the experiments described in this paper are related to the Eastern variant, and the term Serbo-Croatian will hence refer exclusively to it.

2. The Serbo-Croatian language

The term Serbo-Croatian refers to a group of mutually intelligible South Slavic languages (or, alternatively, a South Slavic language with several standard versions) with cca 16 million speakers living across the region of the Western Balkans. Both alphabets it uses (Cyrillic and Latin) are very close to phonemic, making the task of phonetic transcription relatively easy. However, it is a highly inflected language, for which traditional grammars list e.g. seven cases and three genders for nouns and adjectives, as well as seven tenses and three moods for verbs.

The greatest obstacle to the realization of natural-sounding synthetic speech in Serbo-Croatian comes from its complex system of accentuation. Excluding Slovene, Serbo-Croatian is the only Slavic language with a pitch accent, which is a linguistic term of convenience for a variety of restricted tone systems that use variations in pitch to give prominence to a syllable or mora within a word. Unlike languages such as English, German or Italian, which use pitch for pragmatical highlighting, Serbo-Croatian assigns pitch accent at the level of lexicon and uses it to differentiate between word meanings or values of morphological categories. Traditional grammars of Serbo-Croatian define pitch accent through four distinct accent types [10], which involve a rise or fall in pitch associated to either long or short vowels, with optional post-accent lengths, as shown in Table 1.

Table 1. *Traditional Serbo-Croatian accents.*

Name	Examples
short-fall (SF)	kȍcka, mȁk
short-rise (SR)	slobòda, kàfa
long-fall (LF)	rôd, prîča
long-rise (LR)	počétak, mléko
post-acc. length	kapètān, pòžār

Another approach has been suggested in recent linguistic literature, attempting to represent accents as tone sequences. However, studies based on this approach generally differ in the tonal characterization of relevant tone-bearing units. Two of the most prominent such studies, [11] and [12], have been selected as the bases of prosody modeling and prediction in this research.

In [11], which closely follows the ToBI prosody annotation paradigm introduced in [13], the four accents are effectively reduced to two bitonal sequences (falling: H*+L and

rising: L*+H), where the first, starred tone is anchored to the stressed syllable, and the other tone is trailing. Pitch variation is thus realized as the sequence of two tones, and a sparse specification of tones at the surface structure is assumed. Two word-initial boundary tones (%L and %H) are also introduced, in order to account for the behaviour of f_0 at word boundaries, with the latter also being related to focus signaling.

On the other hand, in [12], pitch accent is claimed to be a property of the mora, which is the tone bearing unit, and tones are fully specified at the surface structure. Furthermore, there are also significant differences related to the representation of specific accents as tonal sequences. For example, a trisyllabic word with short vowels carrying the L*+H tonal sequence anchored to the first syllable in [11] is assigned an H-H-L tonal sequence in [12]. Moreover, since in [12] tones are attributed to morae rather than syllables, differences become even more prominent when words with long vowels are considered.

On the phrasal level, earlier analyses of Serbo-Croatian prosody identify three different intonational contours, corresponding to (1) declarative utterances, (2) prosodic questions (with syntax of a declarative statement) and (3) the vocative chant, while in [11] two additional ones have been introduced: (4) signaling of continuation and (5) yes-no question. Based on these assumptions, and according to the ToBI standard, [11] introduces two phrase accents (Ø− and LH−) and three intonational phrase boundary tones (H%, L% and HL%) in order to describe the observed intonational contours.

Most recent analyses of Serbo-Croatian prosody recognize the phonological word and the intonational phrase as the prosodic constituents in Serbo-Croatian. While the pitch range of each phonological word in an intonational phrase is generally smaller than the one of the preceding word, whenever a phrase contains a relatively large number of phonological words (in [11] it is reported to be more than 5), the downstepping trend is interrupted by the adjustment of the pitch range to a higher level in order to allow the speaker to sustain the remaining text within the phrase. While there is some freedom to the choice of the placement of this adjustment, it often coincides with boundaries between major syntactic constituents. The above-mentioned analyses, however, fail to interpret this result as the evidence of a hierarchical prosodic structure within an intonational phrase. The system of prosodic annotation introduced in the following section attempts to account for the variable extent of pitch range adjustment by introducing multiple levels of prosodic phrase boundaries.

3. Overview of the AlfaNumTTS system and the speech database

This section will give a brief description of the AlfaNumTTS speech synthesizer and the speech and language resources it is based on, with special emphasis on the speech database used for the training of the automatic prosody generation module and the conventions used in database annotation.

Most tasks related to high-level synthesis within AlfaNumTTS, including phonetization, part-of-speech (POS) tagging (with the assignment of lexical pitch accent) and the detection of prosodic events such as prosodic phrase boundaries and prominence from text, are carried out by expert systems. Namely, POS tagging relies on a morphological dictionary containing 100,000 lemmas (3.9 million lexical forms), and lexical disambiguation is based on a combination of hand-written grammar rules and transformation rules automatically inferred from a corpus of 200,000 words of previously POS-tagged text [14]. As to the detection of prosodic phrase boundaries and prosodic prominence from text (assignment of specific prosodic markers to POS-tagged text), the current version of the system relies on hand-written rules. Due to the complexity of the general problem of prediction of focus or deaccentuation from POS-tagged text, the system is restricted to assigning prominence markers only in the most obvious cases.

Because of a certain degree of freedom in the choice of positions of both prosodic boundaries and prominence/deaccentuation, no objective assessment of the accuracy of the prosodic marker assignment module is possible, but its positive impact on the naturalness of synthesized speech has been confirmed through subjective evaluation [15]. However, for the purposes of this study prosodic markers have been assigned manually, since it was our intention to evaluate the accuracy of prosody generation at the acoustic level only.

Automatic prediction of prosodic features of speech at the acoustic level (f_0 targets and phonetic segment durations) is based on regression trees trained on a large speech database, which is used for low-level synthesis as well. The database contains 4875 sentences (approximately 4 hours of speech) uttered by a single female voice talent, a professional radio announcer using the ekavian standard pronunciation of Serbo-Croatian. General intonation in the database ranged from neutral to moderately expressive, and the effort was made to keep the speech rate approximately constant. The database was recorded in a sound-proof studio and sampled at 44 kHz.

The database has been annotated with both phonological and phonetic markers (phonemic identity and specific information related to the manner of articulation) as well as markers related to prosody (lexical pitch accents, prosodic phrase boundaries and prosodic prominence). In cases of phones whose articulation consists of more than one phonetically distinct phase (such as occlusion and explosion of stops or vocalic and non-vocalic segments of the vibrant R), annotation was carried out on a sub-phonemic level. Phonemic/phonetic annotation was carried out semi-automatically (AlfaNumASR speech recognition system [16] was used for time-alignment of phone labels, and the results have been inspected and corrected manually), while prosodic annotation was entirely manual. In both cases manual annotation was carried out using AlfaNum SpeechLabel software.

Prosodic annotation included the following:

- marking of the accented vowel of phonological words (in terms of traditional pitch accents) as well as post-accent lengths;
- marking of prosodic phrase boundaries and intonational phrase boundary tones as proposed in [11] (no instance of HL%, the “calling contour”, was found in the database);
- marking of prosodic prominence of particular phonological words (two markers: F+, signaling positive focus of semantically crucial words, and F−, signaling deaccentuation of words representing ‘given’ information or function words originally accented).

Each word in the speech database was also lemmatized and POS tagged, but most of this information is not used in prosody prediction. The inventory of prosodic labels also includes markers reserved for lexical accents, prosodic phrase boundaries and prominence that the labeler was unsure of, preventing debatable phoneme/prosodeme instances from being used for training.

4. Automatic Prosody Generation

F_0 target points and phonetic segment durations, the principal prosodic features necessary for synthesis of natural-sounding speech, are generated using independent regression trees. The database of 4 hours of recorded speech provides approximately 517,000 phoneme instances for training.

4.1. Experiments

In the absence of a specific acoustic model of intonation for Serbo-Croatian, f_0 targets are predicted directly. The prosody generation module predicts the values of f_0 at 1/6, 3/6 and 5/6 of the predicted duration of each voiced phone, first time derivatives of f_0 at the same locations, as well as the difference in f_0 between mid-points of successive vowels. At no point is the entire f_0 curve generated, and the interpolation between f_0 targets is in fact postponed to the low-level synthesis stage, where segments from the speech database are selected so as to fit the f_0 targets with minimum need for postprocessing. The prediction of phone durations is based on an independent regression tree operating on a set of features largely overlapping with the one used for f_0 target prediction. As the size of the database allows the use of an independent regression tree for each phoneme, the phoneme identity is determined within the root node of the duration predictor. The standard approach of predicting *z*-scores instead of actual phone durations is used [17].

Three individual experiments were carried out, with feature sets based on (E1) traditional four accent types, (E2) tonal sequences as proposed in [11] and (E3) tonal sequences as proposed in [12]. An overview of the most important features used is given in Table 2. For the purpose of the experiment, the accent labels in the database were automatically converted to appropriate tone sequences. The phrase accents proposed in [11] were not used in the experiment, as too few instances of HL− were encountered in the database, which contains quite little highly expressive speech. Practically all utterances in it carry the Ø− accent, typical of declarative phrases, *wh*-questions and non-emphatic yes/no questions. The transitional value of tone in E3 is related to the cases when at mid-point of a long (bimoraic) vowel the tone changes.

4.2. Results

Prosody features for 20 utterances withheld from the training data were generated and the resulting synthesized speech was subject to objective and perceptual evaluation. The performance of the prosody predictor was evaluated by objective measures RMSE (root mean square error), MAE (mean absolute error) and CC (correlation coefficient). The objective measures were calculated by the comparison of the actual values from the speech database and the values present in the voiced frames of the synthesized speech. The values of f_0 in the speech database were estimated using PRAAT, without any preprocessing that could have affected the objective measurement. Due to the mismatch in the actual durations of phonetic segments in the database and in the synthesized speech, the objective measures related to f_0 were calculated based on the difference between the values of f_0 at 1/6, 3/6 and 5/6 of the segment in the database and the synthesized speech.

The perceptual evaluation was carried out through listening tests where 20 listeners (native speakers) rated the TTS system performance in terms of naturalness of synthesized speech on a scale from 1 (unnatural, robotic speech) to 5 (speech with apparently natural prosodic features). The listeners were presented with three unmarked sound files per utterance, with prosody features generated according to E1, E2 and E3, in random order. The three versions of the utterance *"A visok je deset centimetara."* (*"And it is 10 centimetres high."*) are given in sound files E1.wav, E2.wav and E3.wav as an example.

It can be noted that the objective measurements related to phone durations have virtually identical values in all experiments, which is not surprising because the features which vary across the experiments are principally related to f_0 prediction.

Table 2. *Feature sets used in the training of regression trees.*

Feature	Values
General prediction features	
lexical stress	stressed, unstressed
phrase boundary	weak, medium, strong, very strong, end of utterance
focus	F+ (prominence), F− (deaccentuation), none
position	position in relation to syllable (onset, nucleus, coda) and to cons./vowel clusters, distance to stressed syllable/word boundary/phrase boundary, measured in the num. of phones/morae/syllables/feet
other	num. of syllables in word, num. of syllables/words in foot/int. phrase, syllable structure, is function/content word
i. p. b. tones	H%, L%
Features specific to duration prediction	
identity	the identity of a particular phone (43 values, including occlusions and releases of stops, several allophones)
manner of art.	vowel, fricative, lateral, stop,...
place of art.	bilabial, alveolar, palatal, velar,...
voicing	voiced, unvoiced
other	the same features for the preceding and the following context; the match or mismatch in feature values between the phone and its prec./foll. context
Features specific to f_0 prediction in Experiment 1	
accent	accent of the vowel/phonol. word (SF, SR, LF, LR, post-acc. length, none)
position	distance to specific accent measured in the number of syllables, accent of the previous/next foot
Features specific to f_0 prediction in Experiment 2	
tone	H, L, none
length	short, long
position	distance to H/L tone measured in the number of syllables, tonal sequence of the previous/next foot (HL, LH)
Features specific to f_0 prediction in Experiment 3	
tone	H, L, transition
length	short, long
position	distance to H/L tone measured in the number of morae, tonal sequence of the previous/next foot (HL, LH)

Table 3. *Results of predictor evaluation.*

	measure	E1	E2	E3
f_0	RMSE [Hz]	18.63	18.24	19.51
	MAE [Hz]	16.33	15.97	17.48
	CC	0.611	0.631	0.605
phone durations	RMSE [ms]	15.85	16.11	16.00
	MAE [ms]	12.02	12.26	12.19
	CC	0.912	0.910	0.911
overall naturalness	MOS	3.94	3.91	3.81

The results are somewhat better than those reported in most recent studies (cf. e.g. [5] for an informative overview), which may be explained by the following factors:

- the size of the database
- the absence of highly expressive content from the database
- the concerted effort to sustain a constant speech rate throughout the recording sessions
- the meticulous procedure of manual inspection of the alignment of phone labels, with precisely defined criteria for label positioning
- the use of markers intended for explicit prevention of the use of phones of apparently abnormal durations (due e.g. to speaker hesitation) in training; the markers were also used to exclude occasional bursts of particularly fast speech from training

Unlike the case of durations, the results of the objective evaluation of f_0 prediction do not compare favorably with state-of-the-art for some other languages (cf. e.g. [3] and [4]). There is practically no difference between the values of the objective measures of the three models, as was the case with durations, with the exception of the model examined in E3 (using tonal characterization given in [12]), where slightly inferior results were obtained. This difference is consistent with slightly lower marks in the perceptive evaluation of this model. It is, however, difficult to say to what extent this result is speaker-dependent having in mind current research methodology and available speech resources. It is our impression that the tone assignment system described in [11] is slightly better suited to modern Serbo-Croatian, and thus a better match to the contents of the speech database, as opposed to [12], which may be more typical of a traditional manner of speaking. However, the fact that the performances of all three models are very similar (in terms of both objective and perceptual evaluation) confirms the initial assumption that traditional accents can be modeled as tone sequences, allowing for the use of a standard prosody annotation scheme such as ToBI.

5. Conclusions and Outlook

The paper presents an integrated approach to fully automatic generation of prosodic features of synthesized speech in Serbo-Croatian, a language for which there exist mutually conflicting theories related to acoustic realization of pitch accent. Both f_0 targets and phonetic segment durations have been predicted using data-driven regression trees, and the results obtained are comparable to those reported in literature for other languages.

The results confirm well-known theoretical principles, such as phrase-final lengthening or the inverse dependency of phone duration from the number of syllables/words in the intonational phrase. Furthermore, the results of the f_0 target prediction according to tonal sequences confirm that traditional accent types of Serbo-Croatian can be successfully modeled as tone sequences, with a slight preference given to [11] over [12].

An important but often underestimated factor that may have a negative influence on prosody prediction accuracy is the speech rate. Namely, it is impossible for the speaker to sustain a constant speech rate regardless of the concentrated effort, especially if the database recording is carried out in multiple sessions. This is particularly harmful in view of the fact that speech rate variability affects both segment durations and f_0 target realization in a non-linear fashion. To the best of our knowledge, this issue has not been addressed in the related research in a systematic way, but it is, nevertheless, easily identified as the source of a considerable amount of variability in the leaves of the regression trees. It is for this reason that our future research will be oriented on robust estimation of speech rate and its incorporation into the CART model.

6. Acknowledgements

The research presented in the paper is supported by the technological project "The Development of Dialogue Systems for Serbian and Other South Slavic Languages" (TR32035) of the Ministry of Science and Technological Development of the Republic of Serbia.

7. References

[1] Sečujski, M., Delić, V., Pekar, D., Obradović, R. and Knežević, D., "An Overview of the AlfaNum Text-to-Speech Synthesis System", Proc. of SPECOM 2007, 3-7 (Add. Vol.), Moscow, Russia, 2007.

[2] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J., "Classification and Regression Trees", Chapman&Hall/CRC, Boca Raton, FL, 1984.

[3] Dusterhoff, K. E., Black, A. W. and Taylor, P., "Using Decision Trees Within the Tilt Intonation Model to Predict F0 Contours", Proc. of EUROSPEECH'99, 1627-1630, Budapest, Hungary, 1999.

[4] Ni, J., Sakai, S., Shimizu, T. and Nakamura, S., "CART-Based Modelling of Chinese Tonal Patterns with a Functional Model Tracing the Fundamental Frequency Trajectories", Proc. of ICASSP, 4253-4256, Taipei, Taiwan, 2009.

[5] Romportl, J. and Kala, J., "Prosody Modelling in Czech Text-to-Speech Synthesis", Proc. of 6th ISCA Workshop on Speech Synthesis, 200-205, Bonn, Germany, 2007.

[6] Wagner, A. and Klessa K., "F0 Contour and Segmental Duration Modeling Using Prosodic Features", Proc. of Speech Prosody 2010, Chicago, IL, 2010.

[7] Lee, S. and Oh, Y.-H., "Tree-Based Modeling of Prosodic Phrasing and Segmental Duration for Korean TTS Systems", Speech Comm., 28(4):283-300, 1999.

[8] Öztürk, Ö. and Çiloğlu, T., "Segmental Duration Modeling in Turkish", Proc. of INTERSPEECH'06, 2378-2381, Pittsburgh, PA, 2006.

[9] Lazaridis, A., Zervas, P., Fakotakis, N. and Kokkinakis, G., "A CART approach for Duration Modeling of Greek Phonemes", Proc. of SPECOM, 287–292, Moscow, 2007.

[10] Ivić, P. and Lehiste, I., "Word and Sentence Prosody in Serbo-Croatian", MIT Press, 1986.

[11] Gođevac, S., "Transcribing Serbo-Croatian Intonation", in S.-A. Jun [Ed], Prosodic Typology: The Phonology of Intonation and Phrasing, 146-171, Oxford Linguistics, 2005.

[12] Inkelas, S. and Zec, D., "Serbo-Croatian Pitch Accent: the Interaction of Tone, Stress, and Intonation", Language, 64(2): 227-248, Linguistic Society of America, 1988.

[13] Beckman, M. and Hirschberg, J., "The ToBI Annotation Conventions", Ohio State Univ., 1994.

[14] Sečujski, M., "Development of Language Resources for the Serbian Language Required for Part-of-Speech Tagging", in S. T. Jovičić and M. Sovilj [Ed], Speech and Language: Interdisciplinary Research III, 125-139, LAAC and IEFPS, Belgrade, 2009.

[15] Stanić Molcer P., Delić, V. and Sečujski, M., "Possibilities of Efficient Evaluation of a TTS System", Proc. of DOGS 2010, 81-84, Iriški venac, Serbia, 2010.

[16] Delić, V., Sečujski, M., Jakovljević, N., Janev, M., Obradović, R. and Pekar, D., "Speech Technologies for Serbian and Kindred South Slavic Languages", in N. Shabtai [Ed], Advances in Speech Recognition, 141-164, InTech, 2010.

[17] Black, A. and Taylor P., "The Festival Speech Synthesis System: System Documentation", Tech. Rep. HCRC/TR-83, University of Edinburgh, UK, 1997.