http://dx.doi.org/10.5755/j01.eee.20.3.4090

Tree-based Phone Duration Modelling of the Serbian Language

S. Sovilj-Nikic[1], V. Delic[1], I. Sovilj-Nikic[1], M. Markovic[2]
[1]Faculty of Technical Sciences, University of Novi Sad,
Trg Dositeja Obradovica 6, 21000 Novi Sad, Serbia
[2]Faculty of Philosophy, University of Novi Sad,
Dr Zorana Dindica 2, 21000 Novi Sad, Serbia
sandrasn@eunet.rs

***Abstract*—Considering the importance of segmental duration from a perceptive point of view, the possibility of automatic prediction of natural duration of phones is essential for achieving the naturalness of synthesized speech. In this paper phone duration prediction model for the Serbian language using tree-based machine learning approach is presented. A large speech corpus and a feature set of 21 parameters describing phones and their contexts were used for segmental duration prediction. Phone duration modelling is based on attributes such as the current segment identity, preceding and following segment types, manner of articulation (for consonants) and voicing of neighbouring phones, lexical stress, part-of-speech, word length, the position of the segment in the syllable, the position of the syllable in a word, the position of a word in a phrase, phrase break level, etc. These features have been extracted from the large speech database for the Serbian language. The results obtained for the full phoneme set using regression tree, RMSE (root-mean-squared-error) 14.8914 ms, MAE (mean absolute error) 11.1947 ms and correlation coefficient 0.8796 are comparable with those reported in the literature for Czech, Greek, Lithuanian, Korean, Indian languages Hindi and Telugu, Turkish.**

***Index Terms*—Decision trees, machine learning algorithms, speech, speech synthesis.**

I. INTRODUCTION

In natural speech the duration of speech segments depends on the context of speech, where that dependence is very complex and involves many factors [1]. The study of speech timing and impact of different phonological, syntactic, physiological and other factors on the duration of speech segments is very important for both understanding the process of speech production and the development of speech synthesis in order to produce high quality synthetic speech [2], [3]. Therefore, a very important component of a text-to-speech (TTS) system is a specialized module (duration system) whose task is modelling segmental duration in natural speech, taking into account various factors. Identifying the most influential factors is a crucial step in the duration modelling process because selecting inadequate or incomplete sets of factors can lead to large errors of duration prediction. Having in mind the nature of the problem, a set of factors that describe the context of speech is composed of only those features that could be extracted from the text. The importance and effect of certain factors on the duration are directly correlated with a particular language and the sets of influential attributes vary considerably in different languages. The choice of this set requires a high level of linguistic information, as well as a statistical analysis of speech corpora in order to determine the accurate value of phone duration in a given language, having in mind the importance of the duration of speech segments from the perceptual point of view [4], [5].

Models for predicting the duration can be divided into two groups: rule-based models and corpus-based models. One of the most well-known models for predicting duration using rules, and the oldest one, was developed by Dennis Klatt [1]. In this type of model, there is an assumption that each phoneme has an intrinsic duration which is inherently one of the distinctive features of phonemes. The intrinsic duration is assigned to a given segment and it is then modified by applying successive rules. In this type of duration system an occurrence of exceptions usually represents a problem because the rules are such that they most often lead to over-generalization. The main advantage of such models is that they do not require large speech corpora, which was of great importance at the time of their creation, when computational resources needed for generating and analysing large speech corpora were not available as today.

However, by the development of computer technology, corpus-based models are becoming more prevalent. Corpus-based statistical models require a large speech corpus because the modelling is done using a machine learning algorithm on large speech corpora. Depending on machine learning approach applied for phone duration modelling, van Santen [5] distinguishes three types of models:

- linear statistical models,
- models obtained using a neural network and
- models based on decision trees.

The first such model for predicting the duration of speech segments in American English was developed by Riley [6] using the CART (Classification and Regression Trees) technique. The technique developed as a combination of statistics and artificial intelligence has many advantages and

Manuscript received April 15, 2013; accepted February 3, 2014.
The research was funded by the Ministry of Science and Technological Development of the Republic of Serbia, within the project TR32035.

as such today represents one of the most commonly used methods for modelling the duration of speech segments. One of the main advantages of the CART method is the ability to find out structural relationships between the predicted and actual values [7]. This is the reason why the CART method is commonly used in the initial stages of phone duration modelling research. The application of this method for modelling phone duration in different languages has given sufficiently good results. Phone duration modelling using the CART method has been implemented for many languages, including Czech [8], Greek [9]–[11], Lithuanian [12], Mandarin, British English, Vietnamese, Serbo-Croatian [13], Korean [14], Turkish [15], and Indian languages Hindi and Telugu [16]. Taking into consideration the above mentioned facts, the authors have decided to select the CART method for modelling phone duration in the Serbian language.

In this paper the authors present the tree-based modelling of phone duration in the Serbian language. The modelling was carried out using two types of decision trees, model tree and regression tree. The performances of these two models are evaluated by objective measures such as root-mean-squared-error (RMSE), mean absolute error (MAE) and correlation coefficient (CC).

In the Introduction the authors present an overview of different approaches to duration modelling, emphasizing in particular the significance of phone duration modelling in speech synthesis. Section II gives the description of the speech database used for extracting a feature set and modelling phone duration as well as a detailed description of the feature set for the Serbian language. Phone duration modelling process using tree-based algorithms is described in Section III. In Section IV experimental results are shown and discussed. Finally, conclusions and further research directions are given in Section V.

II. FEATURE SET FOR THE SERBIAN LANGUAGE

In the duration modelling process a necessary component of a TTS system, which precedes the module for predicting the duration of a speech segment in a given context, is a module that automatically generates the appropriate feature vector that represents each phoneme in the speech database. The elements of the vector describe the characteristics of the speech segment and the context in which it is located, where the value of each feature is actually one of the possible levels of the factors that influence the duration of the speech segment.

If the speech segment in the database is represented by the corresponding feature vector f , and if factor f_j indicates the presence of lexical accent in a syllable and it takes one of mutually exclusive values from the set {stressed, unstressed}, then the feature vector element has one of the possible values of the factor f_j . The space of all factors product $f_1 \times f_2 \times \cdots \times f_n$ is called the feature space $\mathscr{F}$. However, due to the different phonological and other linguistic limitations, not all combinations of different factor values are allowed in any language. Therefore, the linguistic space is defined only by the feature vectors that really occur in a particular language. It is significantly smaller than the feature space and it represents a subset of feature space. On the other hand, the number of possible linguistic combinations of different factor values in any language is extremely large and the recording of such a speech corpus exceeds reasonable time limitations [2]. Thus, when selecting material for speech database recording a lot of attention is focused on finding material that will contain as many different potential linguistic combinations as possible, in order to achieve the highest possible coverage of linguistic space. Despite all the efforts, the speech corpus is often only a small subset of linguistic space [17].

On the basis of the most influential factors that different authors used for phone duration modelling in different languages [1], [8]–[16], [18] and on the basis of previous studies concerning the effect of various factors on the duration of phones in the Serbian language [19], [20], the factors which will be taken into account when developing a model of phone duration in speech synthesis in the Serbian language were selected. All the factors, which will be mentioned later, have been extracted from the speech database in the Serbian language, recorded for the needs of the existing speech synthesizer [21], and various studies which are conducted for the purpose of its improvement [13]. The aforementioned corpus used in this paper contains approximately 2000 sentences and 16000 words, and the majority of its contents are texts from the daily press, which are typically used for such purposes. The speech database was recorded in a sound proof studio and the voice of a female professional radio announcer was employed. She is a Serbian native speaking the Ekavian standard dialect. The recorded speech was sampled at 88.2 kHz. The recorded material was annotated phonetically and prosodically. The AlfaNumASR speech recognition system [22], [23] was used for temporal alignment on the phonetic level and the correction of phone labels was carried out manually using the AlfaNum TTSLabel software [22]. The prosodic annotation involved marking lexical stress (four stress types and post-accentual length), sentence focus and phrase break levels. The prosodic annotation was conducted manually using the AlfaNum TTSLabel software [22].

Each phoneme in the speech database is presented using the appropriate feature vector describing the given speech segment and the context in which the phoneme occurs. In addition, the relevant factors and their potential values in the Serbian language are given. The factors are classified according to the domain of their impact.

- *Nature of the segment*

segment identity: It can be one of 43 different values, including the five vowels in the Serbian language, two different realizations of schwa /ə/ and 25 consonants. Since stops and affricates are labelled in the database as pairs of semi-phonemes that consist of occlusions and explosions in the case of stops and of occlusions and frictions in the case of affricates, the total number of different consonants is therefore 36. Distinction is made between two different types of semi-phone /ə/ that occurs in speech in situations when the phoneme /r/ is found in the consonant environment, or in the syllabic use [24]. In the syllabic usage, phoneme /r/ may be a syllable nucleus and that realization of the vocalic element /ə/ should differ from that in which /r/ is not the nucleus of a syllable.

segment type: vowel, consonant

manner of articulation (for consonants): stop, fricative, affricate, nasal, lateral, semivowel, trill

place of articulation (for consonants): bilabial, labio-dental, dental, postdental, alveolar, palatal, velar

- *Neighboring segments*

segment type: vowel, consonant, silence

manner of articulation (for consonants): stop, fricative, affricate, nasal, lateral, semivowel, trill

voicing: voiced, voiceless

Place of articulation of the preceding and following consonant is not considered because previous studies have shown that it is not a relevant factor [25].

- *Position of segment in syllable*

syllable initial: yes, no

position in a syllable: onset, nucleus, coda

- *Syllable*

lexical stress: stressed, unstressed

stress type: short-fall, long-fall, short-rise, long-rise, post-accentual length

- *Position of syllable in word*

word initial: yes, no

word final: yes, no

- *Word*

part of speech:

- inflected words: noun, verb, adjective, pronoun, number
- indeclinable words: preposition, adverb, conjunction, particle, exclamation

word length: the number of syllables in phonological word

- *Focus*

focus: particularly highlighted word, relatively unimportant word, neutral word

- *Position of word in phrase*

break level: no break, weak break, medium break, strong break, hesitation break, sentence end break

The different break levels correspond to different perceptually detected breaks, in the cases where the break coincides with the interval of silence, it is possible to distinguish between initial, medial and final position of word in the phrase.

III. Phone Duration Modelling

In this paper the CART (Classification and Regression Trees) based method is presented, used for modelling phone duration in the Serbian language.

The CART method, developed as a combination of statistics and artificial intelligence, has a number of advantages and, as such, is one of the most commonly used methods for the duration modelling of speech segments today. One of the main advantages of the CART algorithm is the ability to validate the developed model, which is usually carried out by evaluating the model performance on the data that were not used in the training phase. Also, the CART algorithm is relatively robust in the case of missing data [7]. It allows easy interpretation and processing of the results, statistically selects the most important features and enables a combination of categorical (e.g. the segment identity) and numerical values (e.g. phone duration) of features.

Modelling speech segments duration using the CART technique involves the use of a regression tree for predicting the duration of a given speech segment which is in the database represented by a corresponding feature vector. The formation of the tree consists of several steps: the formation of the question set and the selection of the best question for splitting in the given node; the selection of stopping criterion in a node, or declaration of a given node as a terminal node (leaf); the prediction of a value in a given node.

The most popular splitting criterion is the mean squared error. Suppose Y is the actual duration for training data X, and then the overall prediction error for a node t can be defined as

$$E(t) = \sum_{\mathbf{X}\in t} \left|Y - d(\mathbf{X})\right|^2, \tag{1}$$

where $d(\mathbf{X})$ is the predictive value of Y.

The next step is the selection of the best question which is equivalent to finding the best split for the instances of the node. We should find the question with the largest squared error reduction or the question q^* that maximizes

$$\Delta E_q(t) = E(t) - (E(l) + E(r)), \tag{2}$$

where l and r are the leaves of the node t. We define the expected square error $V(t)$ for a node t as the overall regression error divided by the total number of instances in the node

$$V(t) = E\left(\sum_{\mathbf{X}\in t} \left|Y - d(\mathbf{X})\right|^2\right) = \frac{1}{N(t)} \sum_{\mathbf{X}\in t} \left|Y - d(\mathbf{X})\right|^2. \tag{3}$$

One can notice that $V(t)$ is actually the variance estimate of the duration if $d(\mathbf{X})$ is made to be the average duration of instances in the node. With $V(t)$, we can define the weighted squared error $\overline{V}(t)$ for a node t as follows

$$\overline{V}(t) = V(t)P(t) = \left(\frac{1}{N(t)} \sum_{X\in t} \left|Y - d(X)\right|^2\right) P(t). \tag{4}$$

Finally, the splitting criterion can be rewritten as

$$\Delta \overline{V_t}(q) = \overline{V}(t) - \left(\overline{V}(l) + \overline{V}(r)\right). \tag{5}$$

Regression tree is formed by splitting each node until either of the following conditions is met for a node

1. The greatest variance reduction of the best question falls below a pre-set threshold ~

$$\max_{q\in Q} \Delta \overline{V_t}(q) < \sim. \tag{6}$$

2. The number of instances falling in the leaf node t is below a threshold .

When a node cannot be split further, it is declared a terminal node. The tree building algorithm stops when all nodes are terminal.

Upon the completion of the phase of tree formation by satisfying one of the conditions in (6) a large tree $T_{\max}$ is usually obtained. It can be formed strictly according to the data used in the training phase, but such a tree has no ability to generalize, and it will not have good performance when applied on data that were not used during the training stage. Therefore, it is necessary to find the optimal size of the tree and to avoid data over-fitting. The literature states that there were a number of attempts to overcome this problem, among which Breiman's procedure stands out as the best solution. It consists of several steps:

1. to create the sequence of subtrees

$$T_{\max} \supseteq \dots \supseteq T_k \supseteq \dots \supseteq T_K = t_1, \tag{7}$$

2. for each subtree error rate is estimated,
3. to choose the tree with the lowest error rate, which is the optimal size tree [7].

The procedure described is called cost-complexity pruning. During the formation of a sequence of subtrees produced by pruning some branches the complexity parameter ~ varies from 0 (for $T_{\max}$) up to ∞ (for the subtree containing only the root) so that the following condition is satisfied

$$\min_T \left[\pm^2(T) + \sim \cdot \left| T \right| \right], \tag{8}$$

where: $\pm^2(T)$ is the variance of prediction error for a given subtree and $|T|$ is the number of terminal nodes of a subtree

The prediction of duration of speech segments is done by going through the decision tree, from the root to the leaf of the tree, passing through the internal nodes of the tree by the path which is formed according to the satisfaction of a certain condition on the feature values in each internal node. The leaf contains the predicted value of duration of a given speech segment.

Regression tree is a special case of model tree. The only difference between regression tree and model tree is that for model tree each node contains a linear regression model based on some of the attribute values instead of a constant value. Linear regression model predicts the value for the instances that reach the leaf.

IV. EXPERIMENTAL RESULTS

In this paper duration models have been developed with the M5P (model tree) and M5PR (regression tree) algorithms of WEKA [26]. These algorithms have been used for building binary decision trees on a large speech corpus which contains 98214 phonemes including 38543 vowels and 59671 consonants. SAMPA symbols of phonemes and the number of their occurrences in the speech database for the Serbian language are given in Fig. 1. Phone duration models have been developed for the full phoneme set of the Serbian language as well as for vowels and consonants separately. 10-fold cross-validation procedure has been used to evaluate phone duration models. The evaluation of the duration models developed is carried out using objective measures such as root-mean-squared error (RMSE), correlation coefficient (CC) and mean absolute error (MAE) between the predicted and the actual durations of phones. Prediction performance of each model is also evaluated on unseen (new) data which were not used in the training phase. In this experiment, the whole database was split into two subsets: the training set and the test set. The training set contains 80 % of the database and the remaining 20 % instances constitute the test set.

The root-mean-squared error, mean absolute error and correlation coefficient of both duration models developed using M5P and M5PR algorithms for the full phoneme set are given in Table I. Experimental results shown in this Table are obtained in two different test modes. When testing is performed on new data, which represent 20 % of the whole database (19642 phonemes), the performances of the model are almost the same as in the case of cross-validation test procedure. This is true of both models obtained, indicating a good real prediction performance of the models. One can also notice that the performances of M5PR model are slightly worse than the performance of M5P model.

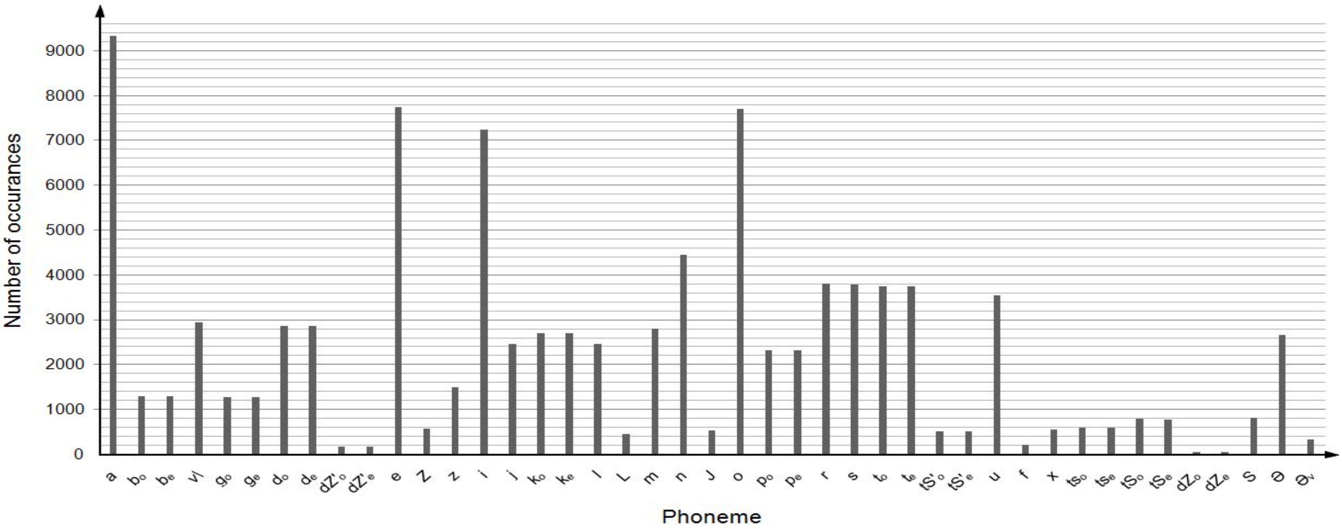

Fig. 1. Phonemes distribution of the Serbian language in the speech database.

This is a very important fact because the application of M5PR model for the prediction of phone duration reduces prediction time even though the number of terminal nodes is larger than in the case of prediction by M5P model, considering the leaves of the tree contain a constant value which is the predicted value of a given phoneme.

TABLE I. PREDICTION PERFORMANCES OF DURATION MODELS FOR THE FULL PHONEME SET (TWO DIFFERENT TEST MODES).

Duration model (test mode)	RMSE [ms]	MAE [ms]	CC
M5P (cross-validation)	15.5579	11.4724	0.9047
M5P (20 % test set)	15.5996	11.4915	0.9050
M5PR (cross-validation)	15.8307	11.6490	0.9012
M5PR (20 % test set)	15.9160	11.7058	0.9008

Table II and Table III show the values of RMSE, MAE and CC for consonant and vowel duration models, respectively. The training set is used for developing the duration models and the evaluation of the model performances is carried out on the test set. The total number of consonants in the training and test sets is 47737 and 11934, respectively. The total number of vowels in the training and test sets is 30835 and 7708, respectively. Performances of these models are comparable with the prediction performance of models obtained for the full phoneme set.

TABLE II. ROOT-MEAN-SQUARED ERROR, MEAN ABSOLUTE ERROR AND CORRELATION COEFFICIENT FOR CONSONANT DURATION MODELS.

Duration model	RMSE [ms]	MAE [ms]	CC
M5P	14.0769	10.3277	0.8788
M5PR	14.4826	10.5714	0.8712

TABLE III. ROOT-MEAN-SQUARED ERROR, MEAN ABSOLUTE ERROR AND CORRELATION COEFFICIENT FOR VOWEL DURATION MODELS.

Duration model	RMSE [ms]	MAE [ms]	CC
M5P	16.6933	12.7239	0.8913
M5PR	17.2721	13.1076	0.8844

In order to improve the model performances obtained the outliers of the speech database have been removed and a new range of phone durations was obtained. This new range of durations contains 96.27 % of the data for the full phoneme set and it was obtained considering the distribution of durations and the number of instances which have extremely small or extremely large durations near the boundary values of the duration range, i.e. around 2 and 290 ms (Fig. 2). Phone duration distribution after removing the outliers is shown in Fig. 3. The distribution of phone durations in the speech database used approximates gamma distribution.

Model performances obtained for the full phoneme set after removing the outliers are given in Table IV. Removing of outliers yields 6.43 % RMSE improvement for regression tree. RMSE, MAE and CC obtained for vowels and consonants after removing the outliers are given in Table V and Table VI.

Table VII shows the percentage of RMSE improvement yielded by removing the outliers. One can notice that the percentage of RMSE improvement is almost the same for both M5P and M5PR developed models as well as for all three different sets. The biggest decrease of RMSE in percentages was obtained for the full phoneme set, whereas the percentage of RMSE decrease for consonants is the smallest.

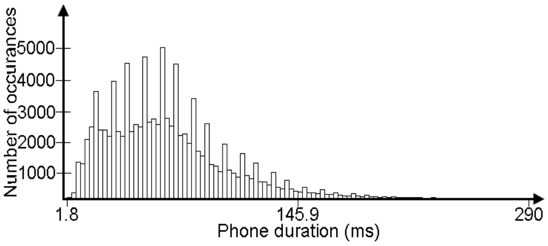

Fig. 2. Phone duration distribution before removing the outliers.

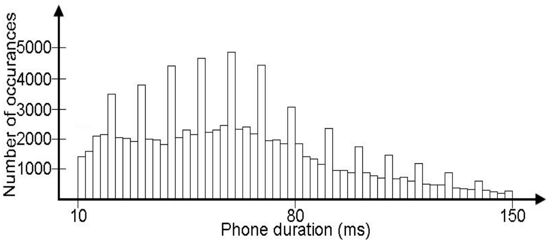

Fig. 3. Phone duration distribution after removing the outliers.

TABLE IV. PREDICTION PERFORMANCES OF DURATION MODELS FOR THE FULL PHONEME SET AFTER REMOVING THE OUTLIERS.

Duration model	RMSE [ms]	MAE [ms]	CC
M5P	14.6028	10.9763	0.8845
M5PR	14.8914	11.1947	0.8796

TABLE V. PREDICTION PERFORMANCES OF CONSONANT DURATION MODELS AFTER REMOVING THE OUTLIERS.

Duration model	RMSE [ms]	MAE [ms]	CC
M5P	13.6274	10.0956	0.8698
M5PR	13.8944	10.2746	0.8643

TABLE VI. PREDICTION PERFORMANCES OF VOWEL DURATION MODELS AFTER REMOVING THE OUTLIERS.

Duration model	RMSE [ms]	MAE [ms]	CC
M5P	15.7466	12.1199	0.8710
M5PR	16.2816	12.5345	0.8615

TABLE VII. PERCENTAGE OF RMSE IMPROVEMENT AFTER REMOVING THE OUTLIERS FOR DIFFERENT SETS.

Set	M5P	M5PR
Phonemes	6.39 %	6.43 %
Consonants	3.19 %	3.12 %
Vowels	5.67 %	5.50 %

TABLE VIII. PREDICTION PERFORMANCES OF DURATION MODELS REPORTED IN LITERATURE FOR DIFFERENT LANGUAGES.

Language	RMSE [ms]	CC
Czech [8]	20.30	0.79
Greek [9]	26.40	0.54
Greek [10]	27.20	0.63
Greek vowels [11]	26.04	-
Greek consonants [11]	29.13	-
Lithuanian vowels [12]	18.30	0.80
Lithuanian consonants [12]	16.70	0.75
Serbo-Croatian [13]	15.85	0.91
Korean [14]	22.00	0.82
Turkish [15]	20.04	0.78
Hindi [16]	27.14	0.75
Telugu [16]	22.86	0.80

Prediction performances of tree-based models for predicting phone durations in different languages reported in the literature are given in Table VIII. It can be noticed that the results achieved using regression tree for the full phoneme set in the Serbian language RMSE 14.8914 ms, MAE 11.1947 ms and CC 0.8796 are comparable with or even outperform the results reported in the literature for different languages.

V. CONCLUSIONS

In this paper tree-based phone duration models for the full phoneme set as well as vowels and consonants of the Serbian language were presented. In the duration modelling procedure a large speech corpus containing 98214 phonemes was used. Removing of outliers was carried out in order to improve model performance. The objective evaluation of the models obtained for the Serbian language was performed and the quantitative measures obtained in terms of RMSE, MAE and CC are comparable with or even outperform those reported in the literature for different languages, including Czech [8], Greek [9]–[11], Lithuanian [12], Serbo-Croatian [13], Korean [14], Turkish [15] and Indian languages Hindi and Telugu [16], developed using regression trees.

In the future, we intend to implement the duration models developed in the speech synthesizer for the Serbian language [21] and to perform subjective evaluation tests of our duration models in order to estimate the quality of synthesized speech on the basis of qualitative measures such as naturalness and intelligibility of speech.

Further research will also include a comparison of duration values predicted by these models with the values obtained by the duration model developed previously for the Serbo-Croatian language [13] and a detailed analysis of differences in terms of influential parameters, concept and complexity among these models.

Considering that speech technologies are directly dependent on the specific language, and so are the most influential factors of segmental duration, as well as the fact that Serbian belongs to the South Slavic language group, the results presented in this study may be used as the basis for modelling duration in other South Slavic languages. Taking into consideration the typological similarities among languages belonging to the same language family, this is the main contribution of this paper, since future research is to be directed towards establishing universal rules regarding the impact of specific factors on the duration of speech segments in South Slavic languages.

REFERENCES

[1] D. H. Klatt, “Linguistic uses of segmental duration in English: acoustic and perceptual evidence”, *Journal of the Acoustical Society of America*, vol. 59, no. 5, pp. 1209–1221, May 1976. [Online]. Available: http://dx.doi.org/10.1121/1.380986

[2] J. P. H. van Santen, “Contextual effects on vowel duration”, *Speech Communication*, vol. 11, no. 6, pp. 513–546, Dec. 1992.

[3] M. Greibus, L. Telksnys, “Segmentation analysis using synthetic speech signals”, *Electronika ir Elektrotechnika (Electronics and Electrical Engineering)*, vol. 18, no. 8, pp. 57–60, 2012.

[4] I. Bulyko, M. Ostendorf, P. Price, “On the relative importance of different prosodic factors for improving speech synthesis”, in *Proc. of ICPhS*, San Francisco, 1999, vol. 1, pp. 81–84.

[5] J. P. H. van Santen, “Timing in text-to-speech systems”, in *Proc. of EUROSPEECH*, Berlin, Germany, 1993, pp. 1397–1404.

[6] M. Riley, “Tree-based modelling of segmental durations”, *Talking Machines: Theories, Models and Designs*. Elsevier, 1992, pp. 265–273.

[7] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984, ch. 8.

[8] R. Batusek, “A duration model for Czech text-to-speech synthesis”, in *Proc. of Speech Prosody*, Aix-en-Provence, France, 2002, pp. 167–170.

[9] A. Lazaridis, P. Zervas, N. Fakotakis, G. Kokkinakis, “A CART approach for duration modeling of Greek phonemes”, in *Proc. of SPECOM*, Moscow, Russia, 2007, pp. 287–292.

[10] A. Lazaridis, V. Bourna, N. Fakotakis, “Comparative evaluation of phone duration models for Greek emotional speech”, *Journal of Computer Science*, vol. 6, no. 3, pp. 341–349, Mar. 2010. [Online]. Available: http://dx.doi.org/10.3844/jcssp.2010.341.349

[11] A. Lazaridis, I. Mporas, T. Ganchev, G. Kokkinakis, “Improving phone duration modeling using support vector regression fusion”, *Speech Communication*, vol. 53, no.1, Jan. 2011. http://dx.doi.org/10.1016/j.specom.2010.07.005

[12] G. Norkevicius, G. Raskinis, “Modeling phone duration of Lithuanian by classification and regression trees, using very large speech corpus”, *Informatica*, vol. 19, no. 2, pp. 271–284, 2008.

[13] M. Secujski, N. Jakovljevic, D. Pekar, “Automatic prosody generation for Serbo-Croatian speech synthesis based on regression trees”, *in Proc. of INTERSPEECH 2011*, Florence, Italy, pp. 3157–3160.

[14] S. Lee, Y. W. Oh, “Tree-based modelling of prosodic phrasing and segmental duration for Korean TTS system”, *Speech Communication*, vol. 28, no. 4, pp. 283–300, Aug. 1999. [Online]. Available: http://dx.doi.org/10.1016/S0167-6393(99)00014-X

[15] O. Ozturk, “Modelling phoneme durations and fundamental frequency contours in Turkish speech”, Ph.D. dissertation, Middle East Technical University, 2005.

[16] N. S. Krishna, H. A. Murthy, “Duration modelling of Indian languages Hindi and Telugu”, in *5-th ISCA Speech Synthesis Workshop*, Pittsburgh, USA, 2004, pp. 197–202.

[17] J. P. H. van Santen, “Assignment of segmental duration in text-to-speech synthesis”, *Computer Speech and Language*, vol. 8, no. 2, pp. 95–128, Apr. 1994. [Online]. Available: http://dx.doi.org/10.1006/csla.1994.1005

[18] J. Pribil, A. Pribilova, J. Matousek, “Comparison of formant features of male and female emotional speech in Czech and Slovak”, *Electronika ir Elektrotechnika (Electronics and Electrical Engineering)*, vol. 19, no. 8, pp. 83–88, Oct. 2013.

[19] S. Sovilj-Nikic, “Trajanje vokala kao jedan od prozodijskih elemenata u sintezi govora na srpskom jeziku”, M.S. thesis, Fakultet tehnickih nauka, Novi Sad, 2007.

[20] M. Markovic, T. Milicev, “The effect of rhythm unit length on the duration of vowels in Serbian”, in *Proc. of 19th ISTAL (Int. Symposium of Theoretical and Applied Linguistics*), Thessaloniki, Greece, 2009, pp. 305–313.

[21] M. Secujski, V. Delic, D. Pekar, R. Obradovic, D. Knezevic, “An overview of the AlfaNum text-to-speech synthesis system”, in *Proc. of SPECOM*, Moscow, Russia, 2007, pp. 3–7.

[22] V. Delic, M. Secujski, N. Jakovljevic, M. Janev, R. Obradovic, D. Pekar, “Speech technologies for Serbian and kindred South Slavic languages”, *Advances in Speech Recognition*. InTech, 2010, pp. 141–164.

[23] V. Delic, M. Bojanic, M. Gnjatovic, M. Secujski, S. T. Jovicic, “Discrimination capability of prosodic and spectral features for emotional speech recognition”, *Electronika ir Elektrotechnika (Electronics and Electrical Engineering*), vol. 18, no. 9, pp. 51–54, Nov. 2012.

[24] S. Guduric, D. Petrovic, “O prirodi glasa [r] u srpskom jeziku”, *Zbornik Matice srpske za filologiju i lingvistiku*, vol. 48, no. 1–2, pp. 135–150, 2005.

[25] T. Crystal, A. House, “Segmental durations in connected speech signals“, *Journal of the Acoustical Society of America*, vol. 83, no. 4, pp. 1553–1573, Apr. 1988. [Online]. Available: http://dx.doi.org/10.1121/1.395911

[26] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, “The WEKA data mining software: An update”, *SIGKDD Explorations*, vol. 11, no. 1, 2009. [Online]. Available: http://dx.doi.org/10.1145/1656274.1656278