ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20171002122930
Original scientific paper

Meta Learning Approach to Phone Duration Modeling

Sandra SOVILJ-NIKIĆ, Ivan SOVILJ-NIKIĆ, Maja MARKOVIĆ

Abstract: One of the essential prerequisites for achieving the naturalness of synthesized speech is the possibility of the automatic prediction of phone duration, due to the high importance of segmental duration in speech perception. In this paper we present a new phone duration prediction model for the Serbian language using meta learning approach. Based on the data obtained from the analysis of a large speech database, we used a feature set of 21 parameters describing phones and their contexts. These include attributes related to the segmental identity, manner of articulation (for consonants), attributes related to phonological context, such as segment types and voicing values of neighboring phones, presence or absence of lexical stress, morphological attributes, such as part-of-speech, and prosodic attributes, such as phonological 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. Phone duration model obtained using meta learning algorithm outperformed the best individual model by approximately 2,0% and 1,7% in terms of the relative reduction of the root-mean-squared error and the mean absolute error, respectively.

Keywords: machine learning; meta learning algorithm; phone duration model; synthesized speech

1 INTRODUCTION

Temporal segmental organization of spoken language is the result of many factors, which are mutually dependent and intricately interconnected [1]. These involve various physiological, phonological, morphologic, syntactic and prosodic factors, and understanding their impact on speech is essential both for understanding the process of speech production and for the development of high quality synthesized speech [2]. A text-to-speech system (TTS) therefore requires a specialized module for segmental duration modeling, which has to take in consideration all the relevant factors.

Two main types of models for predicting segmental duration have been used in TTS systems – rule-based models and corpus-based models.

The oldest, and still rather popular, rule-based model for predicting duration was the one developed by Dennis Klatt [1]. One of the main shortcomings of Klatt's model is that it may lead to over-generalization, especially due to the occurrence of exceptional cases. On the other hand, these models are convenient because they do not require large speech corpora. This was particularly important at the time when they were dominant in speech synthesis, since computational resources needed for generating and analyzing large speech corpora were not available as today.

With the advancement of computer technology, corpus-based statistical models have become more prevalent. These models require a large corpus of spoken language, because the modeling is done using a machine learning algorithm on large corpora. Various machine learning approaches have been applied for phone duration modeling such as artificial neural networks [3, 4] decision trees [5-10], Bayesian models [11], and instance-based algorithms [12].

In this paper the authors present phone duration modeling in the Serbian language using the meta learning algorithm. The modeling was carried out using five different types of individual models and the proposed model. The performances of these models are evaluated by objective measures such as root-mean-squared error (RMSE), mean absolute error (MAE) and correlation coefficient (CC).

This paper is organized as follows: The Introduction section gives an overview of approaches to duration modeling, focusing on the significance of phone duration modeling in speech synthesis. Section 2 provides the description of the speech database used for extracting the set of relevant features and modeling phone duration. It also contains a detailed description of the feature set relevant for the Serbian language. Phone duration modeling process using meta learning algorithm is described in Section 3. The experimental results are presented and discussed in Section 4. Section 5 contains the concluding remarks and proposes further lines of research.

2 FEATURE SET FOR THE SERBIAN LANGUAGE

In order to predict the duration of a speech segment in a given context, a TTS system also requires a module that automatically generates the appropriate feature vector for each phoneme in the speech database. This module precedes the module for predicting the duration of speech segments in the process of speech synthesis.

Most of the factors that influence the duration of segments are universal, and they therefore affect the durational features of segments cross-linguistically. However, some factors may be more marked in some languages than others, so it is important to select the language specific factors when developing a model of phone duration in a speech synthesizer for a particular language. We therefore selected those factors which have been researched in the literature across different languages [1, 6-9], but also the ones found in previous studies concerning the effect of various factors on the duration of phonemes in the Serbian language [13, 14]. These factors have been extracted from the database of spoken Serbian language, recorded for the needs of developing the speech synthesizer for Serbian [15]. This corpus contains approximately 2000 sentences (16 000 words) of read texts taken from daily press, typically used for such purposes. The texts were read by a female professional radio announcer and recorded in a sound

proof studio at 88,2 kHz sampling rate. She is a native speaker of Serbian, who speaks the Ekavian standard dialect. The recorded material was annotated for phonetic and prosodic features. Temporal alignment was done using the AlfaNumASR speech recognition system [16] on the phonetic level, while the correction of phone labels was done manually by means of the AlfaNum TTSLabel software [16]. Prosodic annotation included labels for the four types of lexical stress (long-falling, long-rising, short-falling and short-rising) with additional marking of post-tonic long syllables (post-accentual length). It also involved marking focused elements and phrase break levels. Prosodic annotation was carried out manually using the AlfaNum TTSLabel software [16].

Each phoneme in the speech database is assigned the appropriate feature vector, which contains the information on the segment itself and the context in which it occurs.

The subsequent section of the paper contains the list of relevant factors and their potential values in the Serbian language, classified according to the domain of their impact.

- **Nature of the segment**

segment identity: Serbian has 30 phonemes, 5 vowels and 25 consonants. However, the labeling also had to include two different realizations of the semi-phone schwa /ə/, which is the vocalic element of the consonant /r/. The first type of /ə/ belongs to the phoneme /r/ at syllable margins, and the second when it is part of syllabic /r/ [17]. Stops and affricates are labeled as pairs of semi-phonemes, segmented into the sequences of the occlusion and burst and the occlusion and friction, respectively. The total number of different consonants is therefore 36, and the total of 43 different segment values are accounted for.

segment type: vowel, consonant

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

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

- **Neighboring segments** (the previous and the following segment)

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 [18].

- **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 the 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 break levels were determined on the basis of different perceptually detected relevant discontinuations in the speech chain. In the case of longer intervals of silence (major break), distinction can be made between initial, medial and final position of a word in the prosodic unit.

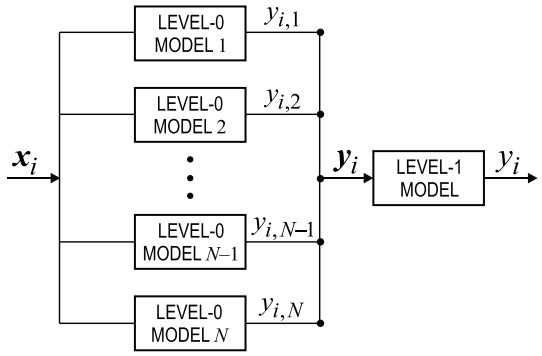
3 PHONE DURATION MODELING USING STACKING ALGORITHM

The basic idea of using meta algorithms is to make more reliable decisions by combining the output of different models [19]. There is no generally accepted preferable way of combining multiple models, so many different variations can be applied. These algorithms often contribute to the improvement of the predictive performance over a single model taking into account the observation that different algorithms perform differently in different situations. In the case of phone duration modeling, different individual phone duration models will produce different errors and the meta learning algorithm, which combines the predictions of individual models in an appropriate way, could compensate for some of these errors. Therefore, meta learning technique will contribute to the increase of the overall phone duration prediction accuracy [19].

Stacked generalization or stacking is one of the meta learning algorithms invented by David Wolpert [19]. It presents a possible way of combing multiple models of different types. Stacking introduces the concept of a metalearner. The metalearner is a learning algorithm which tries to discover how best to combine the output of the base learner or level-0 learner.

General structure of the stacking algorithm is shown in Fig. 1. As one can notice in the figure, the input to the level-1 model consists of the predictions of the level-0 models. A level-1 instance has as many attributes as there are level-0 models, and the attribute values give the predictions of the base learners on the corresponding level-0 instance. The method of cross validation is usually applied for every level-0 learner, ensuring that the level-1 learner uses a full set of training data.

Numerous machine learning methods can be applied for training the level-1 model. Because most of the work is done by the level-0 models, the level-1 learner may be a simple algorithm such as a linear regression or model tree. Different machine learning techniques which are used as a level-0 learner or level-1 learner such as linear regression, model tree, CART (Classification and Regression Trees) and REP (Reduced Error Pruning) Tree will be described briefly in the following paragraphs.



$$\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, ..., x_{i,n}),\ i = 1, 2, ..., M$$

$$\boldsymbol{y}_i = (y_{i,1}, y_{i,2}, ..., y_{i,N-1}, y_{i,N}),\ i = 1, 2, ..., M$$

Figure 1 General structure of stacking algorithm

Linear regression [19] is one of the methods used for the prediction of numerical values. This very simple method is also the oldest method of regression analysis. It has been extensively studied for many years and it is extensively used in many practical applications [20]. The basic idea of this algorithm is that the dependent variable or predictive value is to be presented as a linear combination of attributes which will be taken into account when forming a predictive model and upon which the predictive value depends. Each of the attributes is weighted by the appropriate weighting factor.

Therefore, the dependent variable can be written as:

$$x = w_0 + w_1 a_1 + w_2 a_2 + \ldots + w_k a_k \tag{1}$$

where: x is the predictive value or the phone duration in the case of duration modeling; a_1, a_2,…, a_k are the factors affecting the duration of phonemes; w_0, w_1,…, w_k are weighting factors.

The weighting factors are determined based on the training data, which are in the speech database.

The predicted value of the duration for the first phoneme in the database can be written as:

$$x = w_0 a_0^{(1)} + w_1 a_1^{(1)} + w_2 a_2^{(1)} + \ldots + w_k a_k^{(1)} = \sum_{j=0}^{k} w_j a_j^{(1)} \tag{2}$$

In determining the coefficients w_j, there are $k + 1$ of them, the method of the least squares is applied, and it is necessary to minimize the sum of squares of the differences between the actual and the predicted values over all the training data.

If there are n phones in the database, wherein the i^{th} phoneme denoted with superscript (i) then the sum of the squares of the differences can be represented as:

$$\sum_{i=1}^{n} (x^{(i)} - \sum_{j=0}^{k} w_j a_j^{(i)})^2 \tag{3}$$

where the difference in parenthesis is the difference between the actual and the predicted value of the duration of the i^{th} phoneme in the database. By choosing appropriate coefficients w_j the sum of squares is minimized.

CART method [21] is today probably the most frequently applied method for modeling duration of speech segments in synthesized speech. It 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 takes place in several steps: first. the formation of the question set and selection of the best question for splitting in the given node; selection of stopping criterion in a node, or declaration of a given node as a terminal node (leaf) and the prediction of a value in a given node.

The criterion for splitting is usually the mean squared error. If Y is the actual duration for training data X, then the overall prediction error for a node t can be defined as:

$$E(t) = \sum_{\boldsymbol{X} \in t} |Y - d(\boldsymbol{X})|^2 \tag{4}$$

where $d(\boldsymbol{X})$ 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:

$$\Delta E_q(t) = E(t) - (E(l) + E(r)) \tag{5}$$

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:

$$V(t) = E\left(\sum_{\boldsymbol{X} \in t} |Y - d(\boldsymbol{X})|^2 \right) = \frac{1}{N(t)} \sum_{\boldsymbol{X} \in t} |Y - d(\boldsymbol{X})|^2 \tag{6}$$

One can notice that $V(t)$ is actually the variance estimate of the duration if $d(\boldsymbol{X})$ 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:

$$\overline{V}(t) = V(t)P(t) = \left(\frac{1}{N(t)} \sum_{\boldsymbol{X} \in t} |Y - d(\boldsymbol{X})|^2 \right) P(t) \tag{7}$$

Finally, the splitting criterion can be rewritten as:

$$\Delta \overline{V}_t(q) = \overline{V}(t) - \left(\overline{V}(l) + \overline{V}(r) \right) \tag{8}$$

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 α:

$$\max_{q \in Q} \Delta \overline{V}_t(q) < \alpha \tag{9}$$

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

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

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.

In addition to regression and model trees in the process of duration modeling another algorithm based on decision trees could be used. This is the REP (Reduced Error Pruning) Trees algorithm [22] developed in order to obtain the optimal tree, which means finding a minimum tree size while achieving a minimum error. In this algorithm different sets of data are also used for forming and pruning the tree, wherein the mutual ratio of the amount of data of these two sets is one of the parameters of the algorithm.

4 EXPERIMENTAL RESULTS AND DISCUSSION

In this paper duration models have been developed with LR (linear regression), M5P (model tree), M5PR (regression tree) and REPTree with and without pruning, and STACK (stacking) algorithms of WEKA [23]. These algorithms were used for training the models on a large speech corpus containing 98 214 phones, including 59 671 consonants and 38 543 vowels. Fig. 2 shows the SAMPA (Speech Assessment Methods Phonetic Alphabet) symbols of phonemes and phonemic segments from the Serbian speech database analyzed and the number of their occurrences.

Prediction performance of each model was evaluated in the experiment including unseen (new) data, which were not used in the training phase. The procedure involved splitting the whole database into two subsets: the training set and the test set. The training set included 80% of the database, while the remaining 20% cases were used as the test set. The evaluation of the duration models was performed by means of objective measures, including root-mean-squared error (RMSE), correlation coefficient (CC) and mean absolute error (MAE) between the predicted and actual durations of phones.

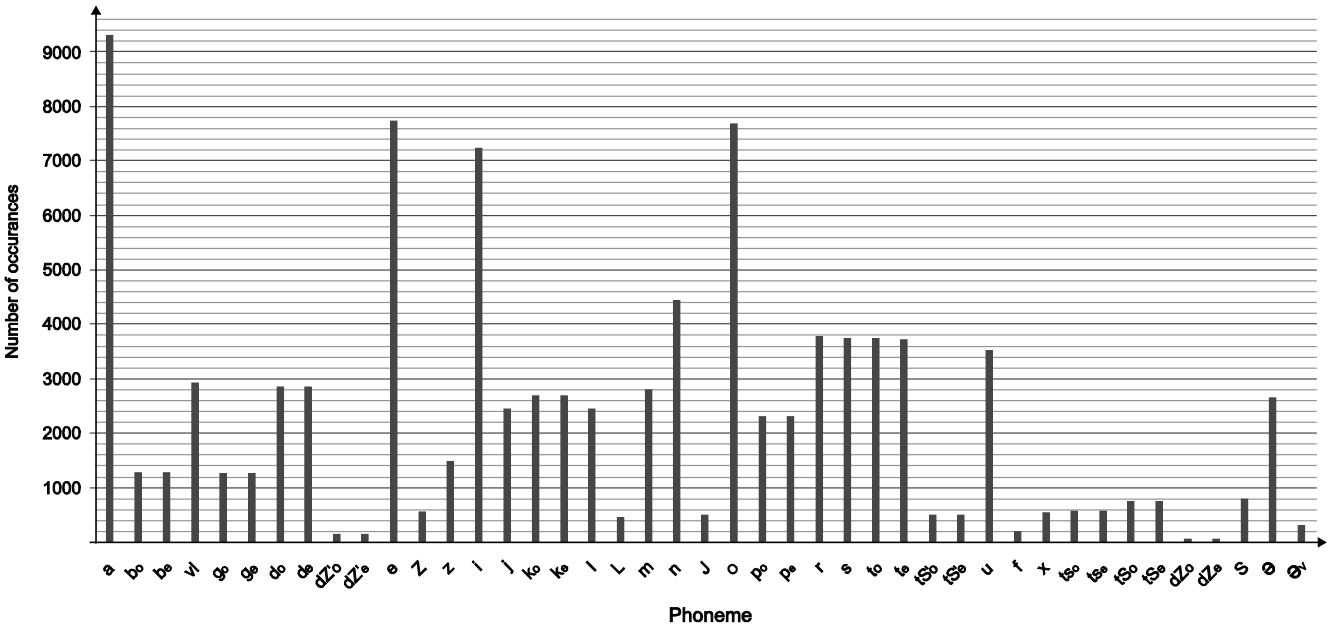


Figure 2 SAMPA symbols of the Serbian phonemes and their distribution in the speech corpus

Table 1 Prediction performances of duration models

Duration model	RMSE / ms	MAE / ms	CC
LR	20,8554	15,7226	0,8224
M5P	15,5996	11,4915	0,9050
M5PR	15,9160	11,7058	0,9008
REPTree no prun.	16,7069	12,1994	0,8901
REPTree	16,8716	12,3388	0,8878
STACK M5P	15,3170	11,3226	0,9085

The root-mean-squared error, mean absolute error and correlation coefficient of duration models developed using LR, M5P, M5PR, REPTree with and without pruning algorithms as well as meta learning STACK algorithm for the full phoneme set are given in Tab. 1. LR, M5P, M5PR, and REPTree with and without pruning were used as level-0 models in stacking algorithm and the M5P has been chosen to be metalearner.

Based on the results presented in Tab. 1 it can be seen that the performances of STACK M5P model are better than the performances of the individual models used as level-0 models in the training phase of STACK M5P model in terms of RMSE, MAE and CC. Among the individual models M5P model has the best prediction performances [24]. This is the reason why M5P was chosen to be used as level-1 model. One can also notice that by the application of linear regression the worst prediction performances model is obtained.

To further improve the achieved STACK M5P model performances, the outliers of the speech database were removed, resulting in a new range of phone durations, which contains 96,27% of the data of the full segment set. It was obtained with regards to the distribution of durations after removing the instances of phones with extremely small or extremely large durations near the

boundary values of the full duration range, i.e. around 2 and 290 ms (Fig. 3).

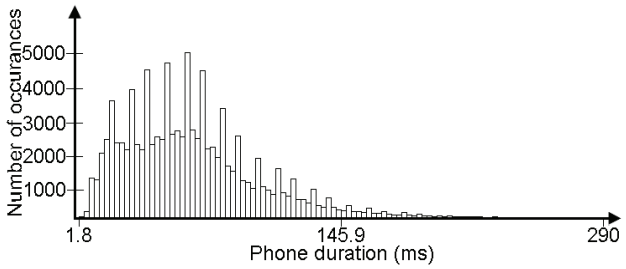


Figure 3 Phone duration distribution before removing the outliers

Phone duration distribution after removing the outliers is shown in Fig. 4. The distribution of phone durations in the speech database used approximates gamma distribution.

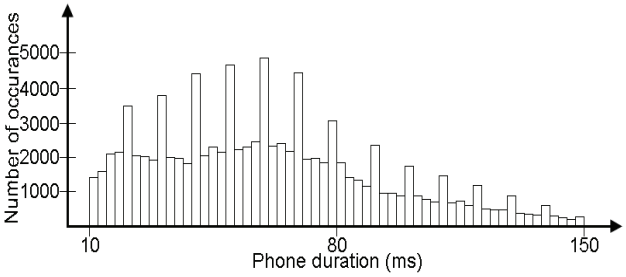


Figure 4 Phone duration distribution after removing the outliers

Model performances obtained for the full phoneme set after removing the outliers are given in Tab. 2.

Table 2 Prediction performances of duration models after removing the outliers

Duration model	RMSE / ms	MAE / ms	CC
LR	18,0529	13,9174	0,8170
M5P	14,6028	10,9763	0,8845
M5PR	14,8914	11,1947	0,8796
REPTree no prun.	15,4042	11,5572	0,8706
REPTree	15,6526	11,7379	0,8660
STACK M5P	14,3118	10,7948	0,8894

Tab. 3 presents the percentage of RMSE improvement achieved after removing the outliers for different models. Obviously, the percentage of RMSE improvement is the highest for the LR model, which is the weakest-performing individual model. On the other hand, the percentage of RMSE decrease is the smallest for the M5P model, which is individual model with the best prediction performances.

Table 3 Percentage of RMSE improvement after removing the outliers for different models

Model	RMSE / %
LR	13,44
M5P	6,39
M5PR	6,43
REPTree no pruning	7,78
REPTree	7,22
STACK M5P	6,56

Fig. 5 is the graph illustration of the percentage of RMSE decrease following the removal of the outliers for different models.

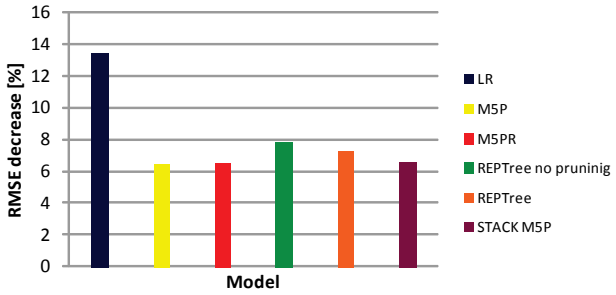


Figure 5 RMSE decrease after removing the outliers

After removing the outliers STACK M5P model outperforms the best individual model M5P by approximately 2,0% and 1,7% in terms of relative reduction of RMSE and MAE respectively.

5 CONCLUSION

In this paper we presented the results of a new phone duration model using meta learning algorithm for the full phoneme set of the Serbian language. The corpus used for the duration modeling procedure contained the total of 98214 phonemes. In order to improve model performance, outliers were removed from the analysis. The model obtained for the Serbian language was subjected to objective evaluation. It was found that the quantitative measures obtained in terms of RMSE, MAE and CC outperform individual models. Therefore, we can conclude that use of the meta learning algorithm contributes to the increase of phone duration prediction accuracy.

Future research should include subjective evaluation of our duration model once it is implemented into the speech synthesizer for the Serbian language [15]. The goal of such a study would be to evaluate the quality of synthesized speech and determine whether and to what extent the proposed model contributes to the naturalness, intelligibility and comprehensibility of synthesized speech.

ACKNOWLEDGEMENTS

This research was funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia, within the projects III 43008, TR 35015 and TR 32035, and it is also the result of the cooperation within CEEPUS project CIII- RO-0058-07-1415 supported by Secretary of Science and Technological Development of the Autonomous Province of Vojvodina.

6 REFERENCES

[1] Klatt, D. H. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of the Acoustical Society of America, 59*(5), 1209-1221. https://doi.org/10.1121/1.380986

[2] Van Santen, J. P. H. (1992). Contextual effects on vowel duration. *Speech Communication, 11*(6), 513-546. https://doi.org/10.1016/0167-6393(92)90027-5

[3] Campbell, W. N. (1992). Multi-level speech timing control, *PhD dissertation*, University of Sussex.

[4] Rao, K. S. & Yegnanaravana, B. (2007). Modeling durations of syllables using neural networks. *Computer Speech and Language, 21*(2), 282-295. https://doi.org/10.1016/j.csl.2006.06.003
[5] Riley, M. (1992). Tree-based modeling of segmental durations. *Talking Machines: Theories, Models and Designs* / Elsevier, 265-273.
[6] Batušek, M. A. (2002). Duration model for Czech text-to-speech synthesis. *Proceedings of Speech Prosody* / Aix-en-Provence, France, 167-170.
[7] Lazaridis, A., Zervas, P., Fakotakis, N., & Kokkonakis G. A. (2007). CART approach for duration modeling of Greek phonemes. *Proceedings of SPECOM* / Moscow, Russia, 287-292.
[8] Öztürk, Ö. (2005). Modeling phoneme durations and fundamental frequency contours in Turkish speech. *PhD dissertation*, Middle East Technical University.
[9] Norkevičius, G. & Raškinis, G. (2008). Modeling phone duration of Lithuanian by classification and regression trees, using very large speech corpus. *Informatica, 19*(2), 271-284.
[10] Sečujski, M., Jakovljević, N., & Pekar, D. (2011). Automatic prosody generation for Serbo-Croatian speech synthesis based on regression trees. *Proceedings of INTERSPEECH 2011* / Florence, Italy, 3157-3160.
[11] Goubanova, O. & King, S. (2008). Bayesian networks for phone duration prediction. *Speech Communication, 50*(4), 301-311. https://doi.org/10.1016/j.specom.2007.10.002
[12] Lazaridis, A., Bourna, V., & Fakotakis, N. (2010). Comparative evaluation of phone duration models for Greek emotional speech. *Journal of Computer Science, 6*(3), 341-349. https://doi.org/10.3844/jcssp.2010.341.349
[13] Sovilj-Nikić, S. (2007). Trajanje vokala kao jedan od prozodijskih elemenata u sintezi govora na srpskom jeziku. *M.Sc. thesis*, Fakultet tehničkih nauka, Novi Sad.
[14] Marković, M. & Milićev, T. (2009). The effect of rhythm unit length on the duration of vowels in Serbian. *Proceedings of 19th ISTAL (International Symposium of Theoretical and Applied Linguistics)* / Thessaloniki, Greece, 305-313.
[15] Sečujski, M., Delić, V., Pekar, D., Obradović, R., & Knežević, D. (2007). An overview of the AlfaNum text-to-speech synthesis system. *Proceedings of SPECOM* / Moscow, Russia, 3-7.
[16] Delić, V., Sečujski, M., Jakovljević, N., Janev, M., Obradović, R., & Pekar, D. (2010). Speech technologies for Serbian and kindred South Slavic languages. *Advances in Speech Recognition* / InTech, 141-164.
[17] Petrović, D. & Gudurić, S. (2010). *Fonologija srpskog jezika*, Beograd: Institut za srpski jezik SANU, Beogradska knjiga i Matica srpska.
[18] Crystal, T. & House, A. (1988). Segmental durations in connected speech signals. *Journal of the Acoustical Society of America, 83*(4), 1553-1573. https://doi.org/10.1121/1.395911
[19] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*, San Francisco, CA: Morgan Kauffman Publishing, Fourth Edition.
[20] Ćosić, P., Lisjak, D., & Antolić, D. (2011). Regression analysis and neural networks as methods for production time estimation. *Tehnički vjesnik/Technical Gazette, 18*(4), 479-484.
[21] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.
[22] Kaariainen, M. & Malinen, T. (2004). Selective Rademacher penalization and Reduced Error Pruning of decision trees. *Journal of Machine Learning Research, 5*, 1107-1126.
[23] Frank, E., Hall, M. A., & Witten, I. H. (2016). *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*, San Francisco, CA: Morgan Kauffman Publishing, Fourth Edition.
[24] Sovilj-Nikić, S., Delić, V., Sovilj-Nikić, I., & Marković, M. (2014). Tree-based phone duration modeling of the Serbian language. *Elektronika i Elektrotehnika (Electronics and Electrical Engineering), 20*(3), 77-82. https://doi.org/10.5755/j01.eee.20.3.4090

Contact information:

Sandra SOVILJ-NIKIĆ, D.Sc. Eng.
Iritel a.d. Beograd
Batajnički put 23, 11080 Beorgad, Serbia
E-mail: sandrasn@eunet.rs

Ivan SOVILJ-NIKI, PhD student
University of Novi Sad
Faculty of Technical Sciences
Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia
E-mail: diomed17@gmail.com

Maja MARKOVIĆ, D.Sc., Associate Professor
University of Novi Sad
Faculty of Philosophy
Dr Zorana Đinđića 2, 21000 Novi Sad, Serbia
E-mail: majamarkovic@ff.uns.ac.rs