INFORMATICA, 2008, Vol. 19, No. 2, 271–284

Modeling Phone Duration of Lithuanian by Classification and Regression Trees, using Very Large Speech Corpus

Giedrius NORKEVIČIUS, Gailius RAŠKINIS

Department of Applied Informatics, Vytautas Magnus University
Vileikos 8, LT-44404 Kaunas, Lithuania
e-mail: g.norkevicius@if.vdu.lt, g.raskinis@if.vdu.lt

Received: May 2007

Abstract. Classification and regression tree approach was used in this research to model phone duration of Lithuanian. 300 thousand samples of vowels and 400 thousand samples of consonants extracted from VDU-AB20 corpus were used in experimental part of research. Set of 15 parameters characterizing phone and its context were selected for duration prediction. The most significant of them were: identifier (ID) of phone being predicted, adjacent phones IDs and number of phones in syllable. Models were built using two different data sets: one speaker and 20 speakers. The influence of cost complexity pruning and different values of pre pruning were investigated. Prediction by average leaf duration vs. prediction by median leaf duration was also compared. Investigation of most vivid errors was performed, speech rate normalization and trivial noise reduction were applied and influence on models evaluation parameters discussed. The achieved results, correlation 0.8 and 0.75 respectively for vowels and consonants, and $RMSE$ of $\sim$ 18 ms are comparable with those reported for Check, Hindi and Telugu, Korean.

Keywords: duration model, phone duration, classification and regression trees, CART, speech rate variability.

1. Introduction

Phone duration model and knowledge about how factors and factor interactions influence phone duration is very important for Language Technology applications like text-to-speech synthesis, automatic speech recognition, and automatic speech signal segmentation. Phone durations are highly context dependent in natural speech. For instance, duration of the phone /a"/ (used SAMPA notation (Raškinis *et al.*, 2003)) is as short as 30 ms in a word „*savo*" ("*my*") and as long as 115 ms in a word „*atvertė*" ("*opened*"). Therefore, the goal of this research is to create a model capable to predict phone durations of Lithuanian on the basis of the context information.

Many types of methods have been recently used to model phone duration, the most common of them are: rule-based (Klatt, 1979), sum-of-products (Santen, 1993; Santen, 1997) or decision tree-based methods (Batušek, 2002; Krishna and Murthy, 2004).

The dominant aproach of ruled based duration modeling method is proposed by D. Klatt. It states that:

1) each phone or group of phones has inherent ($INHDUR$) duration,
2) each of defined rule (value of context dependent parameter $PRCNT$) tries to describe a relative increase or decrease in the duration of a phone,
3) phones cannot be shorter than a certain minimum duration $MINDUR$.

The model is summarized by the formula:

$$DUR = \big((INHDUR - MINDUR) * PRCNT\big)/100 + MINDUR. \tag{1}$$

Thus starting from some intrinsic value, the duration of a phone is modified by successively applying a number of rules. This type of duration model was quite successfully applied to Swedish, English and French languages, however rule-based models often tend to over generalize and cannot handle exceptions well without getting exceedingly complicated (Krishna and Murthy, 2004).

Another popular duration modeling technique is the sums-of-products model, developed by Van Santen (1993, 1997). According to this model the phone described by feature vector d has the duration $DUR(d)$:

$$DUR(d) = \sum_{i \in K} \prod_{j \in I_i} S_{i,j}(d_j). \tag{2}$$

Here, K is a set of indices, each corresponding to a product term. I_i is the set of indices of factors occurring in the ith product term. The parameters $S_{i,j}$ are called factor scales. Construction of the sums-of-products model consists of those three steps:

- Build a category tree – Usually category tree is built by linguists from known duration ratios. One leaf covers phones that are under the similar influence of some factors/parameters.
- Construction of the sums-of-products model for each category tree leaf.
- Model parameter estimation.

It is generally accepted that this method is one of the most accurate and is characterized as the method giving the best prediction and the largest correlation between predicted and actual duration. However, building category tree and attaching models to categories need an exhaustive knowledge about factors and how factor interactions influence phone duration. Though there are some phone duration modeling experiments made by Kaunas University of Technology, there are no published corpora-based investigations on Lithuanian phone durations. Thus, this method is not usable yet.

Therefore one of the machine learning methods – classification and regression tree (CART) modeling (Breiman *et al.*, 1984) can be used for Lithuanian by now. CART is a statistical modeling technique used to predict a dependent variable y corresponding to feature vector X. In many researches of phone durations (Santen, 1993; Santen, 1997; Goubanova, 2003) and in general in the literature (Breiman *et al.*, 1984), it is often stated, that using this method does not provide the most accurate results. Nevertheless, because of its ability to discover structural relationships between the predicted and the measured variables, this method is often used in initial research stages.

2. Modeling by Classification and Regression Trees

As most of machine learning approaches, CART also requires training set L in the form $< X_n, y_n >$, were y_n is training instance dependent value corresponding to feature vector X_n, $n = 1, 2, \dots, N$, where N is the number of instances in data set. The tree construction consists of three steps:

1) building a tree,
2) pruning a tree,
3) selecting an optimal subtree.

These three steps are discussed in Sections 3 and 4.

3. Building a Tree

We start with the tree consisting only of a root node t_1, containing all of the cases in training set L. The task is to find the optimal binary split of t_1 into t_L, t_R. For real valued feature i all splits of the form $x^i < \tau$ are tested. Here τ stands for any real number. For categorical feature i splits have the form $x^i \in \Theta$, where Θ is the set of all possible subsets on values of x^i. Once an optimum split of t_1 is found, optimal splits of each of the two descendant nodes t_L, t_R are recursively sought. This procedure is applied iteratively to all descendants until some stopping condition is fulfilled (usually splitting is performed until the size of node is no less than some specified minimal threshold or the decrease of prediction error is observed). Let $R(T)$ denote error estimate of tree T and $R(t)$ – error estimate of node t. Thus, the choice of the optimal/ best split and prognosis inside terminal node depends on $R(T)$. Two estimates were compared in this research.

1. Root mean square error ($RMSE$):

$$R_{RMSE}(T) = \sqrt{\frac{1}{N}\sum_{i=1}^{M} R_{RSE}(t_i)}, \tag{3}$$

where

$$R_{RSE}(t) = \sum_{X_n \in t} \big(y_n - y(t)\big)^2. \tag{4}$$

2. Mean relative error (MRE):

$$R_{MRE}(T) = \frac{1}{N}\sum_{i=1}^{M} R_{RE}(t_i), \tag{5}$$

where

$$R_{RE}(t) = \sum_{X_n \in t} \frac{|y_n - y(t)|}{y_n}. \tag{6}$$

Let S be the set of all possible splits of a node t. More precisely, for any split s of t into t_L, t_R, let

$$\Delta R(s,t) = R(t) - R(t_L) - R(t_R). \tag{7}$$

Then the best split $s^* \in S$ would be

$$s^* = \arg\max_{s \in S} \Delta R(s,t). \tag{8}$$

Thus, a regression tree is formed by iteratively splitting nodes so as to maximize the decrease in $R(T)$. Notice that in the first $(RMSE)$ case value of $y(t)$ which minimizes $R_{RSE}(t)$ is the mean of values y_n falling into node t:

$$y(t) = \bar{y}(t) = \frac{1}{N(t)} \sum_{X_n \in t} y_n, \tag{9}$$

here the sum is over all $y_n$satisfying condition $X_n \in t$, $N(t)$ is the number of objects falling into node t.

In the second (MRE) case value of $y(t)$ which minimizes $R_{RE}(T)$ is the median of all values $y_n$falling into node t, $\nu(t)$ will denote median value.

4. Pruning

The tree obtained using procedure described in Subsection 2.1 (denote $T_{\max}$) is usually too big and the error estimates are overly optimistic. In other words, tree overfits the data. Minimal error-complexity pruning (Breiman *et al.*, 1984) was used in this research to deal with overfitting. The essence of minimal error-complexity pruning lies in iterative procedure cutting the "weakest" subtrees. The weakest subtree is defined as having the least ratio:

$$\frac{R(t) - R(T_t)}{|\widetilde{T}_t| - 1}, \tag{10}$$

here, $R(T_t)$ is the error estimate of subtree starting from node t, $|\widetilde{T}_t|$ is the number of terminal nodes of a subtree T_t. The outcome of this iterative procedure is an ordered sequence of trees $T_{\max} \supseteq \ldots \supseteq T_k \supseteq \ldots \supseteq T_K = t_1$ of decreasing size. Training data set was used to construct initial trees. Validation set was used to select tree from such sequence having the least error estimate.

Also we used pre-pruning in this research. Pre-pruning is applied in tree construction phase. It does not allow creating nodes that have less samples than some predefined number.

5. Modeling Phone Durations of Lithuanian

5.1. *The Data*

Our analysis of phone durations is based on the VDU-AB20 continuous read speech corpus. The corpus has been compiled and annotated by the Center of Computational Linguistics at Vytautas Magnus University. The VDU speech corpus contains recordings of 20 speakers. Each speaker represents about 1 hour of speech. Thus the corpus contains 20 hours of speech in total. The VDU-AB20 corpus is automatically annotated at 6 levels including phone-level, syllable-level, word-level. In the context of this research, the most important of them is phone-level. The database yielded a total of 300000 vowels and 400000 consonants. About 200 different SAMPA-encoded allophonic symbols (Raškinis *et al.*, 2003) were used in this research. SAMPA codes include information about phone stress, palatalization, being a part of diphone and other features. The procedure of text-to-phone transcription is elaborated in (Norkevičius *et al.*, 2005). Speech signal annotation is based on Hidden Markov model (HMM) methodology and has been achieved using HTK toolkit (Young *et al.*, 2000). A discrete grid/grating with time step of 10 ms was used to automatically determine phone boundaries.

5.2. *Features Used to Predict Phones Duration*

The features that are used to predict duration can be grouped to such levels according to the scope of the context:

Phone level. All features on this level are categorical. Possible values are codes of SAMPA notation (Raškinis *et al.*, 2003):

- target phone identity,
- identities of two preceding and two following phones.

Syllable level. All features on this level are numeral. Measurements are given in a number of phones:

- syllable length,
- distance of the target phone from syllable start,
- distance of the target phone to syllable end.

Word level. All features on this level are numeral. Measurements are given in a number of syllables:

- word length,
- distance of the target syllable from word start,
- distance of the target syllable to word end.

Phrase level. All features on this level are numeral. Measurements are given in a number of words:

- phrase length,
- distance of the target word from phrase start,
- distance of the target word from phrase end.

Extra features. All features on this level are boolean:

- target phone is initial phone of the word,
- target phone is final phone of the word.

5.3. *Criteria for Model Evaluation*

Evaluation of the duration models is usually expressed in terms of:

- Model's error estimate. As mentioned in Subsection 2.1, tree construction and choice of the predicted value for a terminal node directly depends on the criteria used to estimate tree error. Usually a root mean squared error ($RMSE$) is used as tree error estimate and mean as the predicted value for a terminal node. Due to the fact that there was no reasonable explanation found of using this criterion, we additionally used the mean relative error as tree error estimate and median as predicted value for a terminal node. Only one of mentioned criteria (henceforth referred to as mean-based and median-based prediction) is selected at tree construction phase at a time. Nevertheless both error criteria were estimated for a constructed tree.
- Correlation between the actual and the predicted durations.

5.4. *Experimental Setup*

10-fold cross-validation procedure has been applied for all experiments. The database was partitioned into a training set 90%, a validation set 5% and a test set 5%. It must be noticed that the data set used in this research was not validated by experts in general. However, word accentuation was validated for all speakers, text-to-phone conversion was not validated and phone-level annotations were partly validated (only the most prominent errors were corrected) for one speaker. On that score 2 different data sets were separated: data set containing validated data of a single speaker – henceforth referred to as experiment "SS" and data set with a validated accentuation of 20 speakers – henceforth referred to as experiment "MS". There were 26940 vowels and 34281consonants in "SS" and 314114 and 409927 vowel and consonant samples in "MS" accordingly. Considering articulation and acoustics, vowels and consonants are two different phone classes, thus separate models (CARTs) for vowels and for consonants were built in all experiments. Two groups of experiments were performed in this research.

1. Baseline experiments. Separate models were built for SS and MS data sets, all 15 features were used to predict phone duration.
   The following model building options were investigated:
   - pre-pruning parameter: 1 (no pre-pruning), 10, 20, and 30 samples in a leaf;
   - prediction value for a terminal node: mean-based and median-based.

32 baseline experiments were run based on different model building options.

2. Supplementary experiments. Analysis of the largest errors in baseline experiments showed that mainly these were caused by:
   - inter-speaker speech rate variability;
   - intra-speaker speech rate variability;
   - noisy data, i.e., inaccurate phone-level annotations.

   All additional experiments were performed on the „SS" data set. These parameters were investigated:
   - intra-speaker speech rate variability normalization (henceforth referred to as SR norm.);
   - noise reduction (henceforth referred to as NR);
   - prediction value for a terminal node: mean-based and median-based.

5.5. *Experimental Results*

The investigation of influence of the pre-pruning showed that the best results were achieved when it was not allowed to create nodes covering less than 10 samples. Post-pruning of trees (Fig. 1) constructed with pre-pruning parameter equal to 10 did not show any significant changes of models error estimates, however it did decrease the amount of nodes about 12 times (Table 4).

Thus, all results are given for trees constructed with the pre-pruning parameter equal to 10 and after post-pruning was applied. As it was expected least $RMSE$s were achieved when mean prediction (Table 1) was used and least MREs were achieved using median prediction (Table 2). Largest correlation (Table 3) between actual and predicted duration was achieved using mean prediction.

Reducing intra-speaker speech rate variability is highly dependent on how speech rate is defined. We defined speech rate as a relative deviation of the average duration of vowel/consonant inside a time frame from the average global vowel/consonant duration for one speaker. Time frame was defined so it contained no less than 10 sec of speech (unless it was the last frame of the record) and was not allowed to split words. Frames were overlapping with time step of 5 sec. Thus, for vowels speech rate would be

$$SRV_i = \frac{EV_i}{EV}, \tag{11}$$

Table 1

Averages of $RMSE$ estimates after post-pruning, using mean and median prediction. Pre-pruning parameter – 10 samples. Confidence interval was calculated for 0.95 confidence level

	Average $RMSE$			
	SS, Vowel	SS, Consonant	MS, Vowel	MS, Consonant
Mean-based	0.0314 ± 0.001	0.0257 ± 0.0005	0.0367 ± 0.0004	0.0315 ± 0.001
Median-based	0.0326 ± 0.001	0.027 ± 0.0011	0.038 ± 0.0007	0.0333 ± 0.0007

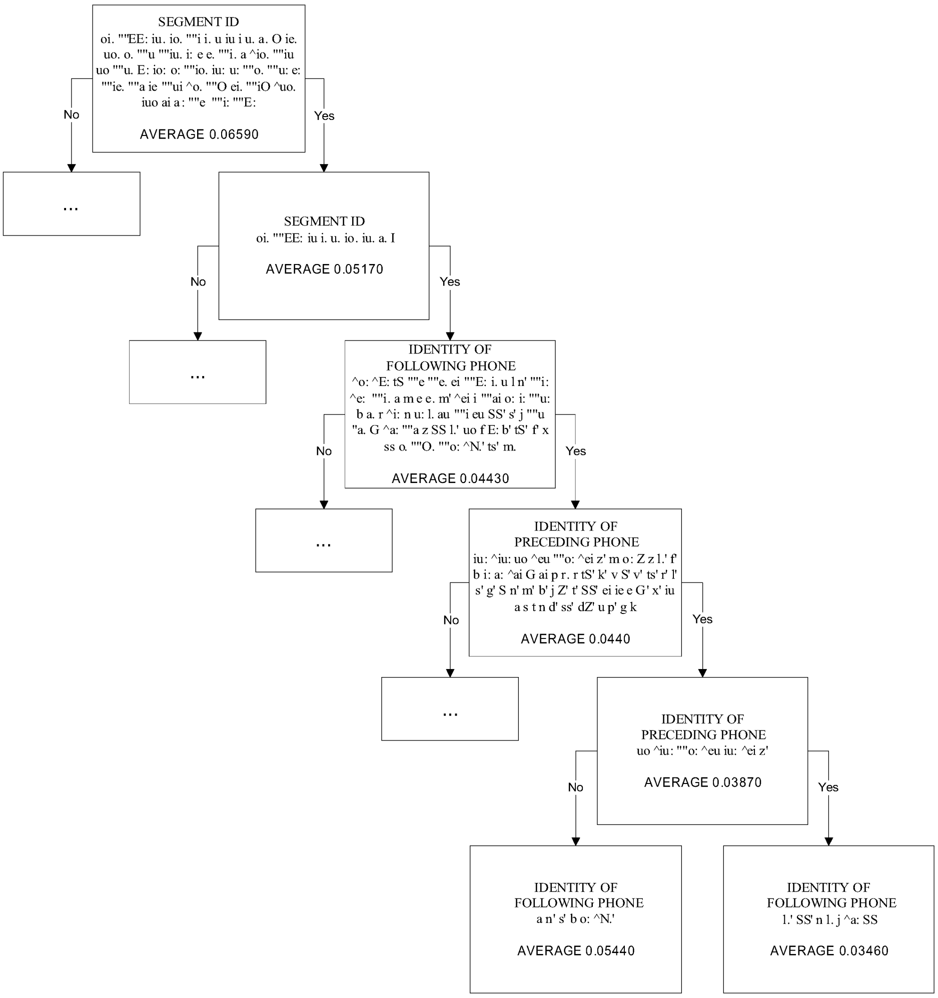
SEGMENT ID
oi. ""EE: iu. io. ""i i. u iu i u. a. O ie. uo. o. ""u ""iu. i: e e. ""i. a ^io. ""iu uo ""u. E: io: o: ""io. iu: u: ""o. ""u: e: ""ie. ""a ie ""ui ^o. ""O ei. ""iO ^uo. iuo ai a : ""e ""i: ""E:
AVERAGE 0.06590
No
Yes
...
SEGMENT ID
oi. ""EE: iu i. u. io. iu. a. I
AVERAGE 0.05170
No
Yes
...
IDENTITY OF FOLLOWING PHONE
^o: ^E: tS ""e ""e. ei ""E: i. u l n' ""i: ^e: ""i. a m e e. m' ^ei i ""ai o: i: ""u: b a. r ^i: n u: l. au ""i eu SS' s' j ""u "a. G ^a: ""a z SS l.' uo f E: b' tS' f' x ss o. ""O. ""o: ^N.' ts' m.
AVERAGE 0.04430
No
Yes
...
IDENTITY OF PRECEDING PHONE
iu: ^iu: uo ^eu ""o: ^ei z' m o: Z z l.' f' b i: a: ^ai G ai p r. r tS' k' v S' v' ts' r' l' s' g' S n' m' b' j Z' t' SS' ei ie e G' x' iu a s t n d' ss' dZ' u p' g k
AVERAGE 0.0440
No
Yes
...
IDENTITY OF PRECEDING PHONE
uo ^iu: ""o: ^eu iu: ^ei z'
AVERAGE 0.03870
No
Yes
IDENTITY OF FOLLOWING PHONE
a n' s' b o: ^N.'
AVERAGE 0.05440
IDENTITY OF FOLLOWING PHONE
l.' SS' n l. j ^a: SS
AVERAGE 0.03460

Fig. 1. Sample of one branch of constructed tree for vowels. Phone duration averages are given for data samples falling in particular node.

Table 2

Averages of MRE estimates after post-pruning, using mean and median prediction. Pre-pruning parameter – 10 samples. Confidence interval was calculated for 0.95 confidence level

	Average MRE			
	SS, Vowel	SS, Consonant	MS, Vowel	MS, Consonant
Mean-based	0.3441 ± 0.014	0.3243 ± 0.0098	0.3131 ± 0.003	0.2772 ± 0.0026
Median-based	0.2904 ± 0.0066	0.2957 ± 0.0099	0.2587 ± 0.0027	0.2478 ± 0.0018

Table 3

Average values of correlation between actual and predicted duration, using mean and median prediction. Pre-pruning parameter – 10 samples. Confidence interval was calculated for 0.95 confidence level

	Average Correlation			
	SS, Vowel	SS, Consonant	MS, Vowel	MS, Consonant
Mean-based	0.7298 ± 0.0162	0.7087 ± 0.0212	0.6508 ± 0.0065	0.5968 ± 0.013
Median-based	0.7171 ± 0.0136	0.6938 ± 0.0270	0.6471 ± 0.0085	0.5698 ± 0.0095

Table 4

Number of terminal nodes after post-pruning, using mean and median prediction. Pre-pruning parameter – 10 samples

	Average number of terminal nodes			
	SS, Vowel	SS, Consonant	MS, Vowel	MS, Consonant
Mean-based	322	573	1253	1647
Median-based	445	249	579	707

here EV_i stands for average vowel duration within a frame i, EV – average vowel duration of the speaker in question. Respectively for consonants speech rate would be

$$SRC_i = \frac{EC_i}{EC}. \tag{12}$$

Normalization of initial data was performed to reduce influence of speech rate (speech rate normalization – SR norm.) in the following way (for vowels):

$$\begin{aligned} D_i = {} & \Big(SRV\big(FR(i)_{left}\big) \times \big(FC\big(FR(i)_{right}\big) - t(i)\big) \\ & + SRV\big(FR(i)_{left}\big) \times \big(t(i) - FC\big(FR(i)_{left}\big)\big)\Big) \\ & \Big/ \Big(FC\big(FR(i)_{right}\big) - FC\big(FR(i)_{left}\big)\Big), \end{aligned} \tag{13}$$

here, $FR(i)$ is a number of frame holding phone i. Indexes $left$ and $right$ are caused by the fact that speech rate frames are overlapping and target phone is influenced by two frames at the same time. FC is a function that gives time value of the frame center, $t(i)$ is the time value of the start of phone i. Similar SR normalization was performed for consonants.

Some experiments were performed with different feature set. It appeared that tree based on the features such as:

- target phone identity,
- identities of preceding and following phones,

- word length in syllables,
- target phone is initial phone of the word,
- target phone is final phone of the word,

lead to best performance. Therefore all successive results are given using the abovementioned feature set and as in previous results trees were constructed with the pre-pruning parameter equal to 10 and after post-pruning was applied. All experiments were run using cross validation procedure described in Subsection 5.4.

For noise reduction (NR) we excluded 10 percent shortest and 10 percent longest samples from the whole data set (SS+MS) each phone type/identification taken separately. SS data set was used for further experimentation. Thus SS data set contained 21212 vowel samples and 25594 consonant samples after applying noise reduction.

As it can be seen, in general both SR normalization and NR gave model improvements, i.e., reduction of error estimates (Table 5, Table 6) and increase in correlation (Table 7). However it has to be noticed that SR normalization was less effective and

Table 5

Averages of $RMSE$ estimates after post-pruning, using mean and median prediction. Pre-pruning parameter – 10 samples. Confidence interval was calculated for 0.95 confidence level

Average $RMSE$				
	Prediction using median		Prediction using mean	
	OS Vowels	OS Consonants	OS Vowels	OS Consonants
Without SR norm.	0.0317 ± 0.0008	0.0266 ± 0.0006	0.0315 ± 0.0008	0.0257 ± 0.0007
SR norm. Applied	0.0311 ± 0.0007	0.025 ± 0.0005	0.0301 ± 0.0007	0.0249 ± 0.0007
NR applied without SR norm.	0.0193 ± 0.0003	0.0173 ± 0.0004	0.0187 ± 0.0004	0.017 ± 0.0003
NR applied SR norm. applied	0.0189 ± 0.0005	0.0172 ± 0.0005	0.0183 ± 0.0003	0.0167 ± 0.0003

Table 6

Averages of MRE estimates after post-pruning, using mean and median prediction. Pre-pruning parameter – 10 samples. Confidence interval was calculated for 0.95 confidence level

Average MRE				
	Prediction using median		Prediction using mean	
	OS Vowels	OS Consonants	OS Vowels	OS Consonants
Without SR norm.	0.2788 ± 0.0051	0.2873 ± 0.0062	0.3326 ± 0.0072	0.3127 ± 0.0095
SR norm. applied	0.2833 ± 0.0104	0.2811 ± 0.0046	0.3219 ± 0.01	0.3111 ± 0.0072
NR applied without SR norm.	0.1869 ± 0.004	0.1738 ± 0.0024	0.1978 ± 0.0046	0.1851 ± 0.0031
NR applied SR norm. applied	0.1885 ± 0.0032	0.1776 ± 0.0046	0.196 ± 0.0029	0.1841 ± 0.004

Table 7

Average values of correlation between actual and predicted duration, using mean and median prediction. Pre-pruning parameter – 10 samples. Confidence interval was calculated for 0.95 confidence level

Average correlation				
	Prediction using median		Prediction using mean	
	OS Vowels	OS Consonants	OS Vowels	OS Consonants
Without SR norm.	0.734 ± 0.0167	0.7033 ± 0.0143	0.726 ± 0.0064	0.7145 ± 0.0142
SR norm. Applied	0.7356 ± 0.0121	0.7121 ± 0.0143	0.7299 ± 0.0122	0.7202 ± 0.0151
NR applied without SR norm.	0.8053 ± 0.0062	0.7547 ± 0.0103	0.8085 ± 0.008	0.7559 ± 0.013
NR applied SR norm. applied	0.8043 ± 0.0083	0.7514 ± 0.0154	0.8051 ± 0.0089	0.7562 ± 0.006

Table 8

Number of terminal nodes after post-pruning, using mean and median prediction. Pre-pruning parameter – 10 samples

Average number of terminal nodes				
	Prediction using median		Prediction using mean	
	OS Vowels	OS Consonants	OS Vowels	OS Consonants
Without SR norm.	376	516	805	1261
SR norm. Applied	593	649	749	1232
NR applied without SR norm.	293	296	606	643
NR applied SR norm. applied	386	233	458	610

hardly observable, but it did reduce the number of terminal nodes (except when NR was not applied and median prediction was used) (Table 8). NR did also reduce the number of terminal nodes, but ir is hard to decide whether it was caused by building a tree on less data samples or because of less noise in the data.

The best achieved results were: Correlations: 0.8 and 0.75 respectively for vowels and consonants; $RMSE$: 0.18 and 0.17 respectively for vowels and consonants; RME: 0.19 and 0.17 respectively for vowels and consonants. Prediction error and correlation coefficients are comparable with those reported for Check (correlation 0.79 , $RMSE$ 20.3 ms) (Batušek, 2002), Hindi (correlation 0.75, $RMSE$ 27.14) and Telugu (correlation 0.8 and $RMSE$ 22.86) (Krishna and Murthy, 2004), Korean (correlation 0.77, $RMSE$ 25.11) (Chung, 2002).

6. Conclusions

A phone duration model of Lithuanian based on the CART technique and built on the VDU-AB20 speech corpus was presented in this paper. Several versions based on different models parameters were presented. Investigation of most noticeable errors was performed, speech rate normalization and trivial noise reduction were applied.

Results of the application of noise reduction points out that the data needs thorough validation. It has to be noted that it is unclear which prediction method - using mean or using median is more advantageous. Taking the abovementioned into account and that it is not always the case that improved statistical model leads to improved speech quality. Also future experiments include inter-speaker normalization and applying different techniques (sum-of-products) to model phone duration.

References

Batusek, R. (2002). A duration model for Czech text-to-speech synthesis. In *In Proceedings of Speech Prosody 2002*. Aix-en-Province, France. pp. 167–170.

Breiman, L., J. Friedman, R. Olshen and C. Stone (1984). *Classification and Regression Trees*. Wadsworth and Brooks.

Chung, H. (2002). Duration models and the perceptual evaluation of spoken Korean. In *In Proceedings of Speech Prosody 2002*. Aix-en-Province, France. pp. 219–222.

Goubanova, O. (2003). Bayesian modelling of vowel segment duration for text-to-speech synthesis using distinctive features. Centre for Speech Technology Research Edinburgh University. `http://www.era.lib.ed.ac.uk/bitstream/1842/1089/2/Goubanova.pdf`

Klatt, D.H. (1979). Synthesis by rule of segmental durations in English sentences. In Lindblom and Ohman (Eds.), *Frontiers of Speech Communication Research*. Academic Press. pp. 287–299.

Krishna, N.S., and H.A. Murthy (2004). Duration modeling of Indian languages Hindi and Telugu. In *5th ISCA Speech Synthesis Workshop*. Pittsburgh, USA. pp. 197–202.

Norkevičius, G., G. Raškinis and A. Kazlauskienė (2005). Knowledge-based grapheme-to-phoneme conversion of Lithuanian words. In *Proceedings of the 10th International Conference on Speech and Computer – Specom.* Patras, Greece. pp. 235–238.

Raškinis, A., G. Raškinis and A. Kazlauskienė (2003). SAMPA (Speech Assessment Methods Phonetic Alphabet) for encoding transcriptions of Lithuanian speech corpora. *Information Technology and Control*, **4**(29), 52–55.

Van Santen, J. (1997). Prosodic modeling in text-to-speech synthesis. In *Proceedings of Eurospeech-97*.

Van Santen, J. (1993) Quantitative modeling of segmental duration. In *Proceedings of the Workshop on Human Language Technology*. pp. 323–328.

Young, S., D. Kershaw, J. Odell, D. Ollason, V. Valtchev and P. Woodland (2000). *The HTK Book*. Entropic, Cambridge.

G. Norkevičius (born in 1980) received his MSc degree in computer science from the Vytautas Magnus University in Kaunas in 2002. Presently, he is a PhD student at the same university. His research interests are speech synthesis and prosodic modeling.

G. Raškinis (born in 1972) received his MSc degree in artificial intelligence and pattern recognition from the University of Pierre et Marie Curie in Paris in 1995. He received doctor's degree in the field of informatics (physical sciences) in 2000. Presently, he works at the Center of Computational Linguistics and teaches at the Department of Applied Informatics of VMU. His research interests include application of machine learning techniques to human language processing.

Lietuvių kalbos garsų trukmės modeliavimas klasifikavimo ir regresijos medžiais, naudojant didelės apimties garsyną

Giedrius NORKEVIČIUS, Gailius RAŠKINIS

Darbe aprašomas eksperimentas, kurio metu lietuvių kalbos garsų trukmės buvo prognozuojamos naudojant klasifikavimo ir regresijos medžius. Tyrimo duomenų bazę sudarė 300 tūkts. balsių ir 400 tūkst. priebalsių pavyzdžių, paimtų iš VDU–AB20 garsyno. Tyrimui naudotas, garsą bei jo kontekstą aprašantis, 15-os požymių rinkinys, svarbiausi jų: prognozuojamo garso identifikatorius, gretimų garsų identifikatoriai, garsų skaičius skiemenyje. Pateikiami eksperimentų rezultatai taikant kvadratinės bei santykinės klaidos medžio mazgų dalinimo kriterijus. Taip pat pateikiami medžio genėjimo minimalios klaidos – sudėtingumo algoritmo pagalba gauti rezultatai bei rezultatai, gauti taikant skirtingus išankstinio genėjimo parametrus. Straipsnyje analizuojamos stambiausios prognozavimo klaidos, aptariami sudarytų kalbėjimo tempo normalizavimo bei duomenų trukšmo mažininimo rezultatai. Sudarytasis modelis leidžia prognozuoti lietuvių kalbos garsų trukmes su $\approx$18 milisekundžių RMSE įverčiu bei 0.8 ir 0.7 koreliacija tarp prognozuotos ir tikrosios reikšmių atitinkamai balsiams ir priebalsiams. Rezultatai palyginami su kitų tyrėjų atliktais (čekų, hindi ir telugų, korėjiečių kalboms) eksperimentais.