INFORMATICA, 2016, Vol. 27, No. 3, 573–586

DOI: http://dx.doi.org/10.15388/Informatica.2016.100

Automatic Parameters Estimation of the D. Klatt Phoneme Duration Model

Pijus KASPARAITIS*, Margarita BENIUŠĖ

Department of Computer Science II, Faculty of Mathematics and Informatics
Vilnius University, Didlaukio 47, LT-08303, Vilnius, Lithuania
e-mail: pkasparaitis@yahoo.com, margarita.beniuse@mif.vu.lt

Received: January 2015; accepted: October 2015

Abstract. Phoneme duration modelling is one of the stages in prosody modelling for text-to-speech systems. The rule-based phoneme duration model proposed by Klatt (1979) is still quite a popular method. One of the main shortcomings of this method is that the values of the parameters are selected in an experimental way. This work proposes a new iterative algorithm for the automatic estimation of the factors for the Klatt model using the corpus of an annotated audio record of the speaker. The phoneme duration models were built for three different Lithuanian speakers. The quality of the estimation of phonemes durations was evaluated by the root mean square error, the mean absolute error and the correlation coefficient.

Key words: text-to-speech synthesis, phoneme duration, Klatt model.

1. Introduction

Phoneme duration modelling is an important part of text-to-speech systems for synthesizing a natural sounding speech. The durations of phonemes may vary widely depending on different speakers and might be completely different from the characteristic average durations estimated by linguists for the language in general. Therefore, one cannot operate with general average durations estimated for the Lithuanian language, and the model of durations should be created specifically for the speaker's voice database (Radziukynienė *et al.*, 2005).

Various methods have been recently used for duration modelling: 1) D. Klatt model (Klatt, 1979), 2) sums-of-products (SoP) (van Santen, 1994), 3) neural networks (NN) (Riedi, 1995), 4) classification and regression tree (CART) (Riley, 1992) and others.

D. Klatt proposed a rule-based model (Klatt, 1979), which was implemented in the MITalk system (Allen *et al.*, 1987). The Klatt duration model was applied to the English, Swedish and French languages. Klatt determined a number of significant contextual effects that influenced the durations of phonemes. Each language has its own specific collection of contextual effects, which influence phoneme durations and should be established under the supervision of linguists. Some examples of the effects, which influence the duration

*Corresponding author.

of a phoneme, are as follows: whether the phoneme is a vowel or a consonant; whether the consonant is voiced or unvoiced; the position of the phoneme in a sentence, the position of the phoneme in a word; the length of the word; the impact of the neighbouring phonemes. For example, the phonemes get longer before the pauses, the consonants get shorter when positioned in a group of consonants, vowels are longer when they are preceded by the voiceless plosives (p, t, k) and so on.

The Klatt model is based on the idea that each phoneme has the minimal duration $D_{\min}$ and inherent duration D_{inh}. The minimal duration depends on the rate, and all minimal durations of phonemes may be scaled for faster or slower rates. The inherent duration is duration of a certain phoneme in a neutral position, but usually the average duration of all instances of a certain phoneme is used. The distance between inherent duration and minimal duration $D_{\text{inh}} - D_{\min}$ may be increased or decreased under the influence of some contextual effects, which mathematically may be expressed as multiplication by corresponding factors (percentages). Consequently the predicted duration of a phoneme is expressed by formula:

$$D = D_{\min} + (D_{\text{inh}} - D_{\min}) f_1 f_2 \dots f_n \tag{1}$$

where $f_1, \dots, f_n$ are factors corresponding to the contextual effects which the phoneme meets and describe how much the duration of the phoneme has changed (stretched or contracted) under the influence of a corresponding effect. One effect may only stretch or contract all the phonemes, none of the effects can stretch one group of phonemes and contract the other group of phonemes.

The sums-of-products duration model (van Santen, 1994) is a general case of the Klatt model. Several product members may be added to the minimal duration in the sums-of-products duration while only one product member is added to the minimal duration in the Klatt model. The sums-of-products duration model is considered to be one of the best models for a phoneme duration prediction. The shortcoming of this model is a lack of methods for an automatic detection of groups of factors: which factors should be multiplied and which should be summarized. The values of the factors may be found by statistical methods, e.g. least-squares regression.

Neural networks seem to give quite reliable prediction results, and the methods for network training are well developed; however, it is almost impossible to interpret the weights of the neural network. The CART method is not as accurate as the sums-of-products duration model or neural networks; however, it is quite simple and easy to interpret.

We chose the rule-based Klatt model (Klatt, 1979) which is a popular method for estimating the phoneme durations. One of the main shortcomings of this method is that factors $f_1, \dots, f_n$ of the model (1) are usually determined manually by the trial and error method. The values of the factors may not be expressed as an analytical solution of Eq. (1) either.

This paper introduces a new iterative algorithm for the automatic evaluation of the Klatt model (1) factors based on the annotated voice recordings of the speaker (Section 2).

This algorithm for an automatic factors evaluation was applied to voice recordings of three different Lithuanian speakers (Section 3). The performance of the phoneme duration

Table 1
Example of contextual effects on the phoneme “a”.

	Sentence:	M	a	m	a	,	m	a	m	a	.
	Duration of phoneme “a” (ms):		222.1		224.8			199.9		353.2	
Effects group Id	Contextual effects:										
gr1	r1 – end of sentence		0		0			0		1	
	r2 – not the end of sentence		1		1			1		0	
gr2	r3 – beginning of sentence		1		0			0		0	
	r4 – not the beginning of sentence		0		1			1		1	
gr3	r5 – end of word		0		1			0		1	
	r6 – not the end of word		1		0			1		0	

prediction model was measured in terms of the root mean square error, the mean absolute error and the correlation coefficient. The list of contextual effects for phonemes was compiled. Contextual effects strongly depend on the language and the list of effects appropriate for one language (e.g. English) cannot be used for other languages. The contextual effects with the greatest and the slightest influence on the durations of the phonemes in Lithuanian were detected.

2. Iterative Algorithm for Calculating the Klatt Model Factors

The model for a phoneme duration prediction is constructed using the corpus, which is generated by annotating an audio record of the text that was read by a speaker. The resulting database is a sequence of names of the phonemes and their durations expressed in milliseconds. The boundaries of the sentences, words and syllables are also marked. Stress marks are included in the names of the phonemes. This kind of the dataset is sufficient to determine which contextual effects do influence the duration of the phoneme.

Let us consider the following example: suppose that our dataset consists of one sentence “Mama, mama”. The durations of the phoneme “a” items are listed in the second row of Table 1. The contextual effects are listed in the second column, the values 1 and 0 denote that a corresponding effect either influences or does not influence the phoneme item, respectively. The contextual effects are organized into groups, e.g. effects related to the position in the sentence or in the word, the number of syllables, etc.

The essence of the proposed algorithm can be briefly described as follows: the contextual effect with the greatest impact on the duration of the phoneme items is identified. The factor of that effect is used to adjust the durations in data. Then in the second step the next contextual effect with the greatest impact is identified: this effect will be either the effect which influences the phoneme items that have not been influenced by the previous effect or the second factor adjusts the durations in the same direction as the first one (e.g. the end of the sentence and the end of the word). In this case the durations will be modified too much. Fortunately, in the third step the first factor can be applied again adjusting the durations in the opposite direction but with a smaller magnitude. The durations are adjusted with a smaller and smaller magnitude in each step and converge to the average durations of the corresponding phonemes. The convergent iterative process is constructed in this way.

Now we shall give a more detailed description of an iterative algorithm for an automatic calculation of the factors for the phoneme duration model (1). This algorithm is repeated for each phoneme separately.

Suppose one phoneme is being examined. Let us denote the ordered list of the durations of the items of this phoneme which are found in the dataset as the vector $\vec{l}^1 = (l_1^1, \ldots, l_M^1)$. The upper index is used for the iteration number. Then the average duration of the items of that phoneme may be calculated as $D_{\text{inh}} = (\sum_{j=1}^{M} l_j^1)/M$ and the minimum of the durations as:

$$D_{\text{MIN}} = \text{MIN}\left\{l_1^1, \ldots, l_M^1\right\}. \tag{2}$$

The value of D_{min} should be chosen from the interval $0 < D_{\text{min}} < D_{\text{MIN}}$ for the algorithm to be able to predict a real minimum duration of the phoneme item if this is the case.

Let's consider the list of contextual effects $r_1, \ldots, r_N$. For each contextual effect r_i, $i = 1, \ldots, N$, there are detected items of the phoneme that are influenced by r_i and the corresponding vector $\vec{r}_i = (r_{i1}, \ldots, r_{iM})$ is constructed, where $r_{ij} = 1$, if the effect r_i does influence the item of the phoneme at the position number j, and $r_{ij} = 0$, if the effect r_i does not influence the item of the phoneme at the position number j. E.g. vectors $\vec{r}_1 = (0, 0, 0, 1)$, $\vec{r}_2 = (1, 1, 1, 0)$ and so on correspond to the data presented in Table 1. Then the number of the items of the phoneme that are influenced by the effect r_i can be calculated as follows: $count_i = \sum_{j=1}^{M} r_{ij}$, $i = 1, \ldots, N$. The calculations described above are performed only once and the values remain unchanged during the iterative process.

The root mean square error and the mean absolute error between two data vectors are calculated respectively as:

$$\text{ErrSQR}\left(\vec{l}, \vec{l}^{new}\right) = \sqrt{\frac{\sum_{j=1}^{M} (l_j - l_j^{new})^2}{M}}, \tag{3}$$

$$\text{ErrABS}\left(\vec{l}, \vec{l}^{new}\right) = \frac{\sum_{j=1}^{M} |l_j - l_j^{new}|}{M}. \tag{4}$$

Next the steps of iteration will be described (the number of iteration is denoted as the superscript $n = 1, 2, \ldots$):

1) For each contextual effect r_i, $i = 1, \ldots, N$, calculate the average duration of the items that are influenced by that effect r_i: $D_i^n = \langle \vec{l}^n; \vec{r}_i \rangle / count_i$, where $\langle \vec{l}^n; \vec{r}_i \rangle$ is the dot product of the vectors.

2) For each contextual effect r_i, $i = 1, \ldots, N$, calculate a corresponding value of the factor f_i^n as follows:

$$f_i^n = \frac{D_i^n - D_{\text{min}}}{D_{\text{inh}} - D_{\text{min}}} \quad \text{or} \quad f_i^n = \frac{\langle \vec{l}^n; \vec{r}_i \rangle - count_i D_{\text{min}}}{count_i (D_{\text{inh}} - D_{\text{min}})}. \tag{5}$$

This formula is derived from (1) where the influence of a single factor is examined.

3) The groups of contextual effects $gr_1, \ldots, gr_G$ are generated with the constraint that one phoneme item may be affected exactly by one effect from the group. Let's construct a vector $\vec{f}_{gr_m}$ for each group $gr_m, m = 1, \ldots, G$, where each coordinate holds the value of that one factor from the group which affects phoneme item in the corresponding position:

$$\vec{f}_{gr_m} = \sum_{r_i \in gr_m} \vec{r}_i f_i^n. \tag{6}$$

E.g. if the group gr_1 consists of effects r_1 and r_2 (see Table 1) then:

$$\vec{f}_{gr_1} = \vec{r}_1 * f_1 + \vec{r}_2 * f_2 = (0, 0, 0, 1)^T f_1 + (1, 1, 1, 0)^T f_2 = (f_2, f_2, f_2, f_1)^T.$$

4) For each group $gr_m, m = 1, \ldots, G$, calculate the predicted durations of the phoneme items $\vec{l}_{gr_m}$ using the factors $\vec{f}_{gr_m}$ of that group as follows:

$$\vec{l}_{gr_m} = D_{\min} + (D_{\text{inh}} - D_{\min}) * \vec{f}_{gr_m}, \quad m = 1, \ldots, G. \tag{7}$$

5) The group of contextual effects whose factors will be used to calculate the durations for the next iteration is selected in this step. Several criteria may be used for selection:

5.1) The group with the factors that produce the least root mean square error (3) between the durations $\vec{l}^n$ and the predicted durations $\vec{l}_{gr_m}$:

$$\text{MinErrSQR} = \text{MIN}\left\{\text{ErrSQR}\left(\vec{l}^n, \vec{l}_{gr_m}\right),\ m = 1, \ldots, G\right\}. \tag{8}$$

5.2) The group with the factors that produce the least mean absolute error (4) between the durations $\vec{l}^n$ and the predicted durations $\vec{l}_{gr_m}$:

$$\text{MinErrABS} = \text{MIN}\left\{\text{ErrABS}\left(\vec{l}^n, \vec{l}_{gr_m}\right),\ m = 1, \ldots, G\right\}. \tag{9}$$

5.3) The group with the sum of factors that are most considerably deviated from value 1:

$$\text{Max}F_1 = \text{MAX}\left\{ \sum_{f_i^n \in gr_m} \left|f_i^n - 1\right|,\ m = 1, \ldots, G \right\}. \tag{10}$$

The situation may occur when the same group of effects is selected repeatedly in sequence. In this case durations are adjusted always in the same direction (positive or negative) and iterative process may not converge to the average duration of the corresponding phoneme. Therefore, it is advisable to restrict the usage of the same group of effects repeatedly step by step.

6) Suppose the group gr_t was chosen in step 5. Then the factors $\vec{f}_{gr_t}$ of that group are used to calculate the durations of the phoneme items, which will be used in the next iteration (starting from step 1), as follows:

$$l_j^{n+1} = D_{\min} + \frac{l_j^n - D_{\min}}{f_{gr_{t_j}}}, \quad j = 1, \ldots, M. \tag{11}$$

These durations converge to the average duration D_{inh} during the iterative process.

The value 1.0 is assigned to the factors of the remaining groups: $f_i^n = 1.0$, if $f_i^n \notin gr_t$.

7) Steps 1–6 are repeated until the completion criterion is met. Several criteria may be used to stop iterations:

7.1) If the difference between MinErrSQR in the adjacent iterations becomes smaller than the predefined threshold;

7.2) If the difference between MinErrABS in the adjacent iterations becomes smaller than the predefined threshold;

7.3) If MaxF_1 becomes smaller than the predefined threshold;

7.4) If the number of iterations exceeds the predefined value.

8) Calculate the final values of the factors $f_1, \ldots, f_N$ accumulated in the iterative process which will be used in formula (1) for predicting the durations:

$$f_i = \prod_{n=1}^{\text{number of iterations}} f_i^n, \quad i = 1, \ldots, N. \tag{12}$$

This iterative algorithm in pseudo code is presented below:

```
Repeat
{
    For each contextual effect r_i
        {
        Calculate the average duration D_i;
        Calculate the factor f_i;
        }
    For each group of contextual effects gr_m
        {
        Calculate the vector f⃗_{gr_m};
        Calculate the predicted durations l⃗_{gr_m};
        }
    Choose the group of effects based on certain criteria;
    Calculate new durations for selected group;
}
Until completion criterion is met;
Calculate accumulated factors
```

The algorithm was tested with a certain amount of synthetic data in order to prove the convergence of the factors calculated in the iterative process to the predefined values f_i which were used to generate those synthetic data.

For example, contextual effects that influence each item of the phoneme “a” in the given sentence were determined in Table 1. The durations were calculated by (1) using the following predefined values: $D_{\min} = 150$, $D_{\text{inh}} = 250$, $f_1 = 1.9$, $f_2 = 0.7$, $f_3 = 1.3$, $f_4 = 0.9$, $f_5 = 1.2$, $f_6 = 0.8$. It was ascertained that the values of the factors calculated by iterative process have converged to the predefined values f_i.

In steps 5.1 and 5.2 the situation may occur when several groups produce the same error, especially when the amount of data is small. In that case, if the algorithm always chooses the same group of effects (for example, the first group from the groups with the same error), the process goes into an infinite loop without any progress in calculations. One way to avoid that kind of loop may be as follows: if method 5.1 produces equal errors for several groups of effects, choose the best group using method 5.2 or vice versa. Method 5.3 does not encounter such problems and converges much faster than 5.1 and 5.2.

The values of thresholds in steps 7.1–7.3 were selected on experimental basis. E.g. the threshold of 0.05 may be used in step 7.3 and the threshold equal to 0.01 may be used in steps 7.1 and 7.2. Experiments showed that the iterative process establishes in about 10–20 iterations in the case of 7.3 criterion and in about 30–40 iterations in the case of 7.1 or 7.2 criteria.

Some modifications can be made to the algorithm. For example, the contextual effects may not be grouped but they can be applied separately in step 5. This method makes the algorithm less complicated; however, the convergence becomes quite slow and the durations calculated by formula (11) do not converge to the average duration D_{inh}.

If there is insufficient data available, it is possible to group similar phonemes (e.g. unvoiced plosive consonants). In this case factors are calculated by applying formula (13) instead of formula (5) where the data of the group of phonemes are used; still the values D_{min} and D_{inh} should remain native for each phoneme in the group:

$$f_i = \frac{\sum_{e\in\text{Group of phonemes}}(\langle \vec{l}^e; \vec{r}_i^e\rangle - count_i^e * D_{\text{min}}^e)}{\sum_{e\in\text{Group of phonemes}} count_i^e (D_{\text{inh}}^e - D_{\text{min}}^e)}. \tag{13}$$

The selection of the value of D_{min} in formula (1) should be discussed in more details. If the value of D_{min} is assumed to be equal to the exact minimal duration D_{MIN} (2), this value can be obtained from (1) if at least one factor f_i is equal to 0. However, every time this zero factor is used, formula (1) will predict the duration to be equal to D_{min}. Consequently, the value D_{min} does not mean the exact minimal duration found in the database. For the iterative algorithm to predict the durations correctly, the value of D_{min} has to be lower than the exact minimal duration, i.e. $D_{\text{min}} = D_{\text{MIN}} - \delta$, where the value of δ is chosen by experiments when the error of the predicted durations is the smallest and the correlation between the actual durations and the predicted ones is the strongest.

3. Results of Phoneme Duration Modelling for the Lithuanian Language

3.1. *Dataset*

Three different phoneme duration models were built on the basis of the corpora of the annotated audio records of one male and two female Lithuanian speakers. The records were annotated by hand using PRAAT software (http://www.fon.hum.uva.nl/praat/). The resulting corpus is the sequence of names of the phonemes and their durations expressed in milliseconds. The boundaries of sentences, words and syllables are also marked.

The stress has a great influence on the phoneme duration (Pakerys, 1982). Phonemes are also divided into long and short ones, if the length is a differential attribute of the phonemes in a corresponding language. In the present work the phonemes were encoded according to the system described in Kasparaitis (2005). That system treats stressed/unstressed and long/short vowels as different phonemes. For example, the phoneme "a" is a short unstressed vowel, "A" is a short stressed vowel, "aa" is a long unstressed vowel, "Aa" and "aA" are long vowels stressed with the falling and rising accent, respectively. Consequently, such features as stressed phonemes, long/short phonemes are not included in the list of the contextual effects since such phonemes are treated as different ones.

Factors of the duration model were calculated for each speaker separately. Each speaker read 5000 short sentences (over 3 hours of audio record, about 160 000 phonemes). The same audio records were used in creating the Lithuanian text-to-speech synthesizer based on a unit selection. For a more detailed description of how a set of sentences was generated see Kasparaitis and Anbinderis (2014).

The records of each speaker were divided into three parts: the training dataset consisted of 3000 sentences (over 96 000 phonemes); the validation and testing datasets consisted of 1000 sentences each (over 32 000 phonemes in each part). The training dataset was used to calculate the parameters of the phoneme duration model (1). The validation dataset was used to select a collection of parameters that gave the best duration prediction (e.g. select the best value of $D_{\min}$). The testing dataset was used to evaluate the errors generated by the model.

3.2. *Experimental Results*

A list of contextual effects that can influence the duration of a phoneme was compiled on the basis of Chung *et al.* (1999), Klatt (1979), Möbius and van Santen (1996). Vowels and consonants were treated as different phoneme classes according to their articulation. Hence, separate contextual effects were created for vowels and consonants. The list of effects on the vowels is presented in Table 2 and that on the consonants – in Table 3. The effects were classified into 13 groups each of which consisted of 2–4 effects. Groups of effects were generated with the restriction that one phoneme could be affected exactly by one effect of the group.

The iterative algorithm described in the previous section was applied to calculate the values of the factors f_i for the Klatt model (1) using datasets of three Lithuanian speakers. Method 5.3 was used to select the best effect group in step 5. The iterative process was stopped when the value (10) in step 7.3 was less than 0.05.

For example, the values of the factors f_i calculated for a duration prediction (1) of a long unstressed vowel "aa" for speaker Female1 are listed in Table 2. There were 663 items of the phoneme "aa" in the training dataset, the average duration $D_{\text{inh}} = 129.14$ ms, the exact minimal duration $D_{\text{MIN}} = 30$ ms, the value of $D_{\min} = 20$ ms was chosen.

For example, the values of the factors f_i obtained for a duration prediction (1) of the consonant "m" for speaker Female1 are listed in Table 3. There were 1836 items of the

Table 2
The values of the Klatt duration model factors f_i for the long unstressed vowel "aa" (Lithuanian speaker Female1).

Effects group Id	Contextual effects on vowels	Factors f_i
gr1 (end of sentence)	r1 End of sentence	1.495
	r2 End of sentence before consonants	1.000
	r3 Not at the end of sentence	0.699
gr2 (beginning of sentence)	r4 Beginning of sentence	1.000
	r5 Beginning of sentence after consonants	1.075
	r6 Not at the beginning of sentence	0.999
gr3 (end of word)	r7 End of word	0.998
	r8 End of word before consonants	0.927
	r9 Not at the end of word	1.004
gr4 (beginning of word)	r10 Beginning of word	1.131
	r11 Beginning of word after consonants	1.155
	r12 Not at the beginning of word	0.985
gr5 (number of syllables)	r13 In monosyllabic word	1.052
	r14 In disyllabic word	0.988
	r15 In 3-syllable word	1.023
	r16 In the word with more than 3 syllables	0.974
gr6 (vowel before vowel/consonant)	r17 Vowel before vowel	1.161
	r18 Vowel before voiced consonant	1.160
	r19 Vowel before unvoiced consonant or pause	0.997
	r20 Vowel before sonorant	0.963

Table 3
The values of the Klatt duration model factors f_i for the consonant "m" (Lithuanian speaker Female1).

Effects group Id	Contextual effects on consonants	Factors f_i
gr7 (end of sentence)	r1 End of sentence	2.592
	r2 Not the end of sentence	0.986
gr8 (beginning of sentence)	r3 Beginning of sentence	1.064
	r4 Not the beginning of sentence	0.998
gr9 (end of word)	r5 End of word	0.803
	r6 Not the end of word	1.008
gr10 (beginning of word)	r7 Beginning of word	1.000
	r8 Not the beginning of word	1.000
gr11 (number of syllables)	r9 In monosyllabic word	0.920
	r10 In disyllabic word	1.014
	r11 In 3-syllable word	0.992
	r12 In the word with more than 3 syllables	1.007
gr12 (consonant in the group of consonants)	r13 In the group of consonants	1.000
	r14 Not in the group of consonants	1.000
gr13 (sonorant before vowel/consonant)	r15 Sonorant before vowel	0.980
	r16 Sonorant before consonant or pause	1.123

phoneme "m" in the training dataset, the average duration of all items was 67.06 ms, the exact minimal duration was 20 ms, the value of $D_{\min} = 10$ ms was chosen.

As explained in the previous section, the value of $D_{\min}$ used in (1) must be smaller than the exact minimal duration D_{MIN} (2) of the phoneme found in the training dataset.

Series of the experiments with different values of $D_{\min}$ were performed. The best value of $D_{\min}$ was selected on the basis of the prediction error/correlation coefficient within the validation dataset. The algorithm of this process in pseudo code is presented below:

```
For each speaker
{
    For each phoneme
    {
        Calculate D_inh of the training dataset;
        Find D_MIN in the training dataset;
        For D_min = D_MIN − 5ms down to 0 with step 5 ms
        {
            Calculate f_i using D_inh, D_min and the training dataset;
            Calculate the prediction error Err_valid and
            the correlation coefficient R in the validation dataset;
        }
        Find the collection {D_inh, D_min, f_i} corresponding
        to the best Err_valid and R;
        Calculate the prediction error Err_test in the testing dataset;
    }
}
```

When the best values of D_{inh}, $D_{\min}$, and $\{f_i\}$ are chosen the root mean square error ErrSQR (3) and the mean absolute error ErrABS (4) between the durations predicted by the Klatt method and the actual durations in the testing dataset were measured (see Tables 4 and 5). These results were compared to the corresponding errors when the values of the predicted durations were taken as the average durations of phonemes. Table 4 shows that

Table 4
The estimation of the root mean square error in the testing dataset between the actual durations (ms) and: 1) the durations predicted by the Klatt method; 2) when the values of the predicted durations are assumed to be average durations of the corresponding phonemes.

Speaker	Root mean square error		
	Klatt method	Average durations method	Improvement
		Vowels	
Female1	29.50	36.44	19.05%
Female2	18.88	20.74	8.98%
Male	19.12	27.74	31.10%
Average	22.50	28.31	19.71%
		Consonants	
Female1	21.26	23.79	10.63%
Female2	14.89	17.37	14.26%
Male	16.94	20.21	16.17%
Average	17.70	20.46	13.69%

Table 5
The estimation of the mean absolute error in the testing dataset between the actual durations (ms) and 1) the durations predicted by the Klatt method; 2) when the values of the predicted durations are assumed to be average durations of the corresponding phonemes.

Speaker	Mean absolute error		
	Klatt method	Average durations method	Improvement
		Vowels	
Female1	21.56	27.57	21.82%
Female2	13.76	15.41	10.76%
Male	14.45	19.65	26.49%
Average	16.59	20.88	19.69%
		Consonants	
Female1	16.33	18.08	9.68%
Female2	11.26	12.78	11.88%
Male	12.83	14.69	12.69%
Average	13.47	15.18	11.42%

Table 6
Values of the correlation coefficient between the actual durations of phonemes in the testing dataset and the durations predicted by the Klatt method.

	Correlation coefficient	
Speaker	Vowels	Consonants
Female1	0.78	0.73
Female2	0.79	0.75
Male	0.84	0.77
Average	0.80	0.75

the average improvement in the root mean square error is on average 19.71% for vowels (reaches even 31.10% for a male speaker) and 13.69% for consonants. Table 5 shows that the average improvement in the mean absolute square error is on average 19.69% for vowels (comes up to 26.49% for a male speaker) and 11.42% for consonants.

The following formula was used to calculate the correlation coefficient R:

$$R = \frac{cov(\vec{l}, \vec{l}^{predicted})}{\sigma \sigma^{predicted}} \tag{14}$$

where $\vec{l}$ is the vector of actual durations of the phonemes, $\vec{l}^{predicted}$ is the vector of the durations predicted by the Klatt method, σ and $\sigma^{predicted}$ are standard deviations of $\vec{l}$ and $\vec{l}^{predicted}$ respectively. The values of the correlation coefficient between the actual durations and the durations predicted by the Klatt method are presented in Table 6 for all the vowels and all the consonants separately.

The following results for the testing dataset were obtained when durations were measured in milliseconds: *ErrSQR* 22.5 for the vowels and 17.70 for the consonants; *ErrABS*

Table 7
Influence of the exclusion of a group of effects on the root mean square error of the predicted durations (*ms*) in the testing dataset (speaker Female2).

Excluded group of contextual effect	The root mean square error
Vowels	
gr1 (end of sentence)	**17.140**
gr2 (beginning of sentence)	16.690
gr3 (end of word)	16.737
gr4 (beginning of word)	16.757
gr5 (number of syllables)	16.699
gr6 (vowel before vowel/consonant)	16.776
Consonants	
gr7 (end of sentence)	**17.305**
gr8 (beginning of sentence)	**16.834**
gr9 (end of word)	16.778
gr10 (beginning of word)	16.707
gr11 (number of syllables)	16.710
gr12 (consonant in the group of consonants)	**16.937**
gr13 (sonorant before vowel/consonant)	16.741

16.59 for the vowels and 13.47 for the consonants; the correlation coefficient *R* 0.8 for the vowels and 0.75 for the consonants. These results are comparable to the results achieved for Lithuanian (*ErrSQR* 18, *R* 0.8) (Norkevičius and Raškinis, 2008), for Check (*ErrSQR* 20.3, *R* 0.79) (Batusek, 2002), Hindi (*ErrSQR* 27.14, *R* 0.75) (Krishna and Murthy, 2004), Korean (*ErrSQR* 25.11, *R* 0.77) (Chung, 2002).

The influence of each group of contextual effects was estimated in order to ascertain which groups were important and which ones had an insignificant influence on the durations of the phonemes in the Lithuanian language. The method of exclusions was used: each group of effects was excluded from the list of the effects one by one and the iterative algorithm described in Section 2 was applied. Prediction errors were estimated. If the exclusion of the chosen group gave a significant increase in the prediction error, the excluded group of the effects was considered to be significant. The influence of the exclusion of the effect group on the prediction error for speaker Female2 is shown in Table 7. This method revealed that the most significant groups of the effects for all three speakers were as follows: the end of the sentence for the vowels (group 1), the end of the sentence for the consonants (group 7), the beginning of the sentence for the consonants (group 8) and a consonant in the group of consonants (group 12).

4. Conclusions

A new iterative algorithm for an automatic parameters evaluation of the Klatt phoneme duration model (1) was introduced in this paper. The Klatt model was proposed in 1979 and was applied to English, Swedish, French and other languages, but the parameters of the model were usually determined manually for each language separately. The proposed algorithm for automatic model parameter calculation may be used for any language. Several

criteria for the selection of the contextual effect group with the greatest impact on duration in each iteration step and several completion criteria were tested. Experiments showed the fastest convergence when the criterion based on the factors deviation from the value 1.0 is chosen compared with the methods based on the prediction error. The algorithm was implemented and used to calculate the Klatt model parameters for the Lithuanian language on the basis of the data of three different speakers. The quality of prediction was evaluated by the root mean square error, the mean absolute error and correlation. These results are comparable with the results obtained by other authors, e.g. for Lithuanian (Norkevičius and Raškinis, 2008) who used classification and regression trees. The proposed algorithm also gave better prediction results as compared with the results obtained when the average duration values were used for prediction. Besides, the investigation showed that contextual effects with the strongest influence on the phoneme duration in Lithuanian appeared to be: the end of the sentence, the beginning of the sentence for the consonants and a consonant in the group of consonants.

Acknowledgements. This research became possible thanks to the project "Services Controlled through Spoken Lithuanian Language (LIEPA)" (No. VP2-3.1-IVPK-12-K-01-001).

References

Allen, J., Hunnicut, S., Klatt, D. (1987). *From Text to Speech: the MITalk System*. Cambridge University Press, Cambridge.

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

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

Chung, H., Huckvale, M., Kim, K. (1999). A new Korean speech synthesis system and temporal model. In: *Proceedings of 16th International Conference on Speech Processing*, vol. 1, pp. 203–208.

Kasparaitis, P. (2005). Diphone databases for Lithuanian text-to-speech synthesis. *Informatica*, 16(2), 193–202.

Kasparaitis, P., Anbinderis, T. (2014). Building text corpus for unit selection synthesis. *Informatica*, 25(4), 551–562.

Klatt, D.H. (1979). Synthesis by rule of segmental durations in English sentences. In: Lindblom, B., Ohman, S. (Eds.), *Frontiers of Speech Communication Research*. Academic Press, New York, pp. 287–300.

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

Möbius, B., van Santen, P.H.J. (1996). Modeling segmental duration in German text-to-speech synthesis. In: *Proceedings of ICSLP*, vol. 4, pp. 2395–2398.

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.

Pakerys, A. (1982). *Lithuanian Language Prosody*. Mokslas, Vilnius (in Lithuanian).

Radziukynienė, I., Šurkutė, A., Kazlauskienė, A., Tamošiūnaitė, M. (2005). Analysis of Lithuanian speech sound lengths and pitch for prosody generation. In: *The Second Baltic Conference on Human Language Technologies*, pp. 323–328.

Riedi, M. (1995). A neural-network-based model of segmental duration for speech synthesis. In: *Proc. Eurospeech*, 1995, pp. 599–602.

Riley, M. (1992). Tree-based modeling of segmental duration. In: Bailly, C.B.G, Sawallis, T.R. (Eds.), *Talking Machines: Theories, Models and Designs*, Elsevier Science Publishers, pp. 265–273.

van Santen, J.P.H. (1994). Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, 8, 95–128.

P. Kasparaitis graduated from Vilnius University (Faculty of Mathematics) in 1991. He became a PhD student at Vilnius University in 1991. In 2001 he defended his PhD thesis. Presently he is an Associate Professor at Vilnius University. Current research includes text-to-speech synthesis and other areas of computer linguistics.

M. Beniušė received her master degree in Vilnius University (Faculty of Mathematics) in 1999, defended her PhD thesis in Vilnius University in 2005. Presently she works at Vilnius University as a lector. Current research interest is text-to-speech synthesis.

Automatinis D. Klatt garsų trukmių modelio parametrų apskaičiavimas

Pijus KASPARAITIS, Margarita BENIUŠĖ

Garsų trukmių modeliavimas yra viena iš balso sintezės iš teksto sudedamųjų dalių. D. Klatt pasiūlytas trukmių modelis – vienas iš seniausių, bet vis dar populiarių metodų. Bene pagrindinis jo trūkumas tas, kad modelio parametrų vertės paprastai apskaičiuojamos rankiniu būdu. Šiame darbe pasiūlytas naujas iteracinis algoritmas, automatiškai apskaičiuojantis Klatt modelio parametrus, naudojant diktoriaus anotuotų įrašų garsyną. Sudaryti trijų lietuvių diktorių garsų trukmių modeliai. Modelių prognozavimo kokybė įvertinta skaičiuojant vidutinę kvadratinę paklaidą, vidutinę absoliutinę paklaidą ir koreliaciją.