A Duration Model for Czech Text-to-Speech Synthesis

Robert Batůšek

Laboratory of Speech and Dialogue, Faculty of Informatics,
Masaryk University, Brno, Czech Republic
xbatusek@fi.muni.cz

Abstract

Duration model is a standard part of current speech synthesizers. Many types of models have been used recently, e.g. multiplicative models ([8]), sum-of-products models ([9]) or decision tree-based models ([4],[5]). This paper follows a decision tree approach. It describes several versions of the duration model for Czech speech synthesis. The model presented here will be implemented in the Czech TTS system Demosthenes ([1],[2]).

1. Introduction

Almost all contemporary text-to-speech synthesizers include a prosody modeling module. Traditionally, at least three prosody components are modeled: segment durations, intonation and intensity. This paper deals with the duration modeling module.

We will use the following notation in the rest of the paper. Speech segment f is represented by an n-dimensional feature vector:

$$f = (f_1, f_2, \ldots, f_n)$$

$d(f)$ will denote the real duration of the segment f and $\bar{d}(f)$ duration computed by a predictor. The task of duration modeling is to predict values $\bar{d}(f)$ as close as possible to values $d(f)$. Let us now briefly summarize the approaches used to model segment durations.

Multiplicative model. This model supposes that the segment duration of the particular segment can be computed as:

$$\bar{d}(f) = F_1(f_1) \times F_2(f_2) \times \cdots \times F_n(f_n),$$

where F_1 represents the intrinsic duration of the segment and $F_2, \ldots, F_n$ represent the effects of other factors.

Sum-of-products model. This model was introduced in [9]. It is a generalization of the multiplicative model (and some other models). According to this model, segment duration is expressed as:

$$\bar{d}(f) = \sum_{i \in T} \prod_{j \in I_i} S_{i,j}(f_j),$$

where $S_{i,j}$ is the function representing the influence of factors i, j. In the multiplicative model, $|T| = 1$ and $I_1 = \{1, 2, \ldots, n\}$.

CART-based model. A CART-based modeling successively divides the feature space to minimize the prediction error. Finally, it constructs a tree representing the partition of the feature space. The CART technique is discussed in more detail in the next section.

Accuracies of several recent duration models are presented in table 1.

2. CART-based Approach

Classfication and regression trees (CART) are a statistical modeling technique used to predict a value of a variable y using the corresponding feature vector f. The prediction process may be illustrated on an example.

Suppose, we have a speech segment f described by the following feature vector (for the description of features see below):
f = (phid=t_S,previd=i,nextid=e, ... ,wordpos=m)
We want to predict the duration of this segment using the tree from figure 5. First, we ask the question in the root node: *Is f long vowel?* As f is not a long vowel, we continue asking the question in the right descendant of the root node: *Is f unvoiced plosive, fricative or affricate?* We continue asking questions until a terminal node is reached. A value in the terminal node is the predicted duration (107.7 ms for our sample segment).

The tree construction consists of three steps: building a tree, pruning subtrees and selecting an optimal tree. To build a tree we need a training (or learning) set L in the form $\{(f^n, y^n); n = 1, 2, \ldots, N\}$, where f^n are feature vectors of corresponding objects and y^n values of the dependent variable. We start with the tree consisting only of a root node t_1 containing all of the cases in L. The task now is to find the optimal binary split of the data. For real-valued feature i all splits of the form $f_i^n < \tau$ are tested. For the M-valued categorical feature i, splits have the form $f_i \in \Theta$, where Θ goes through all subsets of the set of all possible values of the feature i. The best split across all features is selected and the data in the root node is splitted and sent into nodes t_L, t_R. This procedure is applied recursively to all descendants until a stopping condition is fulfilled. Root mean square error is used as a splitting criterion:

$$\sum_{f \in t_L} (d(f) - y_{\bar{L}})^2 + \sum_{f \in t_R} (d(f) - y_{\bar{R}})^2$$

After the tree construction phase, we have a relatively large tree T_{max}. We successively prune some branches and construct a tree sequence $T_{max} \supseteq \cdots \supseteq T_k \supseteq \cdots \supseteq T_K = t_1$. Among these trees we select the best tree using a test sample independent on a training sample.

The CART-based approach has several advantages. Let us mention at least the simplicity of interpretation of the final classifier and the possibility of combination of categorical and real-valued features. For more detailed description of CART the reader may consult [3].

3. Description of the Database

We collected a set of short articles selected from Czech newspapers. A native male speaker read isolated sentences. The database consists of 56 sentences and 5081 phones.

Author	Language	Model Type	Correl. coeff.	RMSE (ms)
Shih	Chinese	multiplicative	—	25
van Santen	Am. English	sum-of-products	0.9	—
Lee	Korean	CART	0.82	22
Chung	Korean	CART	0.73	26

Table 1: A comparison of several duration modeling components.

Utterances were segmented manually using the *Praat* program enabling to display the speech wave, spectrogram and other acoustic characteristics. Phone, syllable, word and phrase boundaries were labelled. Phrase boundaries corresponded to pauses in the waveform. Syllable boundaries were labelled by only one annotator. The exact syllable boundary placement depended only on his subjective opinion.

The data was divided at random into training set (70% — 3,563 segments) and test set (30% — 1,518 segments).

We used a phoneme system with 10 vowels (5 short vowels and 5 corresponding long vowels) and 25 consonants. Pause was marked by an extra symbol. We use the SAMPA alphabet (see [6]) to describe the phoneme identities in this paper.

4. Baseline Model

Based on the literature (see e.g. [4],[5],[8],[9]) and some informal observations the following features were used in the first experiment:

- **phid** — the identity of the current phoneme. This feature is categorical and it has 35 possible values. During the tree construction phase, all subsets of the set of all values should be investigated. This is computationally infeasible for such a large set and we decided to use another solution. Based on the phonetic knowledge, we manually defined subsets to be investigated. Another possible solution is given in [5], but it has not been tested.
- **previd, nextid** — the identity of the preceding and following phoneme. The symbol / stands for pause where neccessary. The same approach was used for defining subsets to be investigated. A set of subsets of the *previd* feature slightly differs from the set of subsets of the *nextid* feature.
- **sl, wl, pl** — syllable, word and phrase length in phones. Ordered numerical features with levels 1–6, 1–20, 1–150. Although it is theoretically possible to have longer words or phrases in Czech, they did not appear in our database.
- **spb, spe** — phone position in syllable from the beginning (end) in phones. Ordered numerical feature with levels 0–6.
- **wpb, wpe** — phone position in word from the beginning (end) in phones. Ordered numerical feature with levels 0–20.
- **wordpos** — word position in a prosodic phrase. Categorical feature with levels {phrase-initial, phrase-middle, phrase-final}. All subsets of this set are investigated, although the subset {phrase-initial, phrase-final} probably does not make sense.

The regression tree was grown and then pruned back using the 0SE rule (see [3]). The results are shown in table 5, row 1. The model is comparable with other CART-based models. Nevertheless, the average error is still quite large. In the rest of the paper we will try to find the ways how to improve it.

5. Feature Revision

When we look at the data in a more detail, we can see that the effects of some features do not correspond to our intuitive expectations. Let us look at figure 5, which shows the average durations of segments depending on the word length. The decreasing trend is obvious, which means that in general phones in long words are shorter than phones in short words. However, this assertion does not hold consistently. For instance, phones in words consisting of 6 phonemes are in average longer than phones in words consisting of 3 phonemes and so on. The same pattern appears at features spb, spe (fig. 5), wpb and wpe.

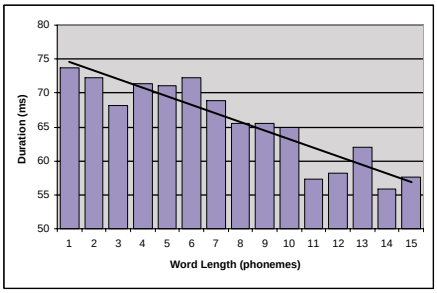

Figure 1: Average durations of phonemes depending on the word length in phones

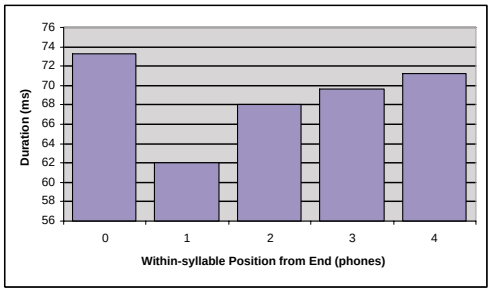

Figure 2: Average durations of phonemes depending on the within-syllable position from the end of the syllable

In fact, although these features are defined as ordered, their ordering does not match the influence on segmental duration. We may provide two solutions how to overcome this drawback.

1. A feature is no longer handled as ordered, but as categorical with the set of possible values $\{1, 2, \ldots, M\}$. We may either go through all the subsets of this set,

or we can sort the set according to the influence of each factor value on segmental duration and ask only $M - 1$ questions. For instance, for the feature spe we would ask whether $spe \in \{0\}, spe \in \{0, 4\}, spe \in \{0, 4, 3\}, spe \in \{0, 4, 3, 2\}$ or $spe \in \{0, 4, 3, 2, 1\}$.

2. We will analyse the data and find the feature analogous to the original feature with the monotonous influence on the segmental duration. The feature *word length in syllables* (instead of word length in phones) is an example of such feature (see figure 5).

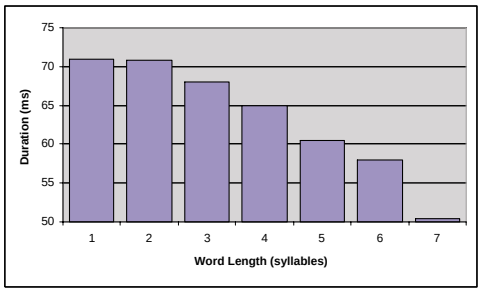

Figure 3: Average durations of phonemes depending on the word length in syllables

All features used in the first experiment were analysed and their influence on the segmental duration investigated. The second experiment was run with the following features:

- **phid, previd, nextid, sl, pl, wordpos** — The same as in the previous experiment.
- **wsl** — Word length in syllables. Ordered numerical feature with levels 0–10.
- **wpsb** — Syllable position in word from the beginning. Categorical feature with values {word-initial syllable, not word-initial syllable}.
- **wpse** — Syllable position in word from the end. Categorical feature with values {word-final syllable, not word-final syllable}.
- **ppsb** — Syllable position in the phrase from the beginning. Categorical feature with values {phrase-initial syllable, phrase second syllable, other syllable}.
- **ppse** — Syllable position in the phrase from the end. Categorical feature with values {phrase-final syllable, phrase penultimate syllable, other syllable}.
- **sp** — Phoneme position in the syllable. Categorical feature with values {syllable onset, syllable nucleus, syllable coda}. This feature makes sense only for consonants. Although most of them appear in onsets and codas, Czech consonants *r, l* or *m* may be also syllabic and thus form a syllable nucleus.

Row 2 of table 5 shows the results of the second experiment. The accuracy is slightly higher. What is probably the most suprising is the size of the final tree. With only 13 (!) terminal nodes (i.e. 12 questions) we are able to predict the segment duration with very high precision.

The tree is presented in the figure 5. In each non-terminal node a splitting question and the average duration of the segment is presented. The average duration and root mean square error are depicted in each terminal node.

Model	Tree size	Correl. coeff.	RMSE (ms)
Baseline	85	0.77	22.1
Revised Features	13	0.79	21.1
Combined	—	0.79	20.3

Table 2: RMSEs and correlation coefficients of various versions of CART-based duration model of Czech speech.

6. Phoneme Class-Specific Models

Figure 6 show the influence of the *wsl* feature on the short vowels. Short vowels in unisyllabic words are shorter than in bisyllabic. On the other side, while the influence of the *spe* feature on all phonemes cannot be easily interpreted, its influence on short vowels is clear (see fig. 6). These effects lead us to building a separate tree classifiers for various phoneme classes.

The data were divided into three classes: vowels, sonorant consonants and other consonants. The lack of data didn't allow us to make more subtle partition. For each class, class-specific features were derived. The tree was built for each class separately.

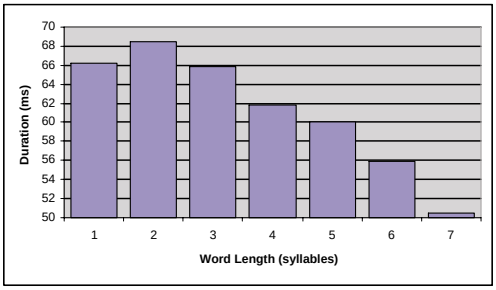

Figure 5: Average durations of short vowels depending on the word length in syllables

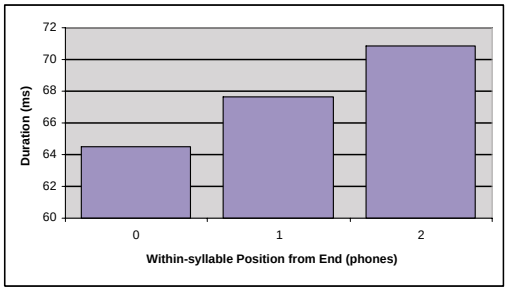

Figure 6: Average durations of short vowels depending on the within-syllable position from the end of the syllable

The third row of table 5 shows the accuracy of the model combining class specific models. Phoneme class-specific model improves error rate, correlation coefficient remains roughly the same.

7. Future Research

It is clear that root mean square error is not the optimal measure of a duration model accuracy. Predicting 20 ms instead of 60 ms is much more serious mistake than predicting 200 ms instead

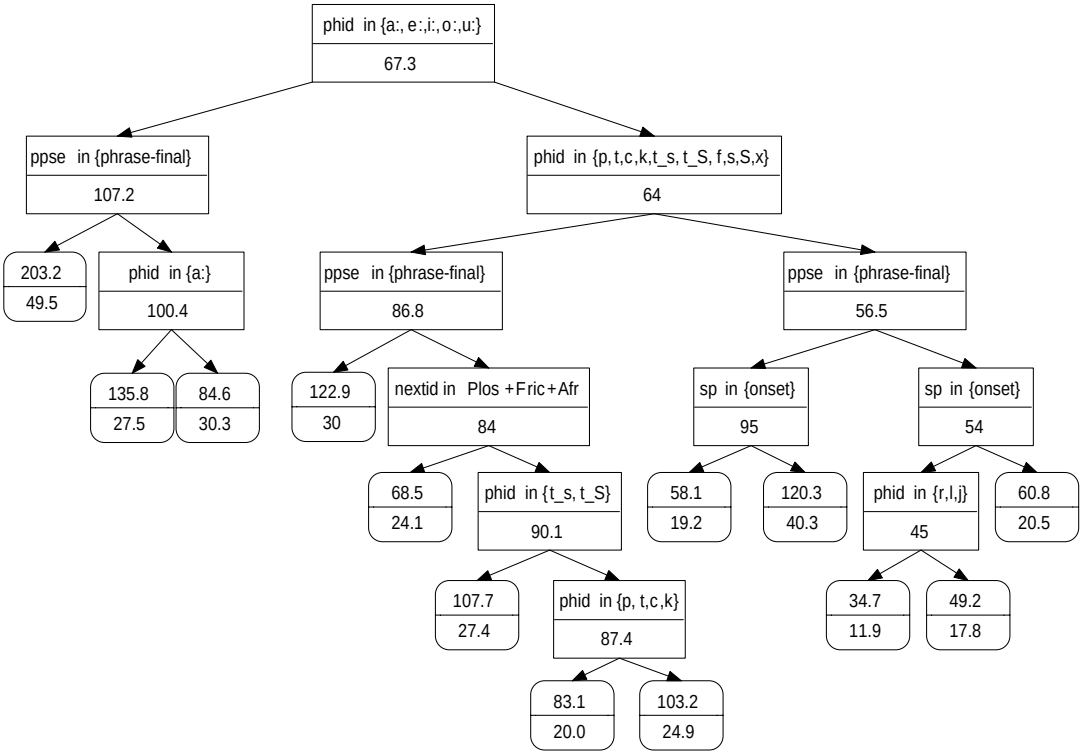

Figure 4: Regression tree for predicting segmental duration.

of 240 ms. It sounds reasonable that the error measure should depend also on the absolute values of both predicted and real durations. Therefore, the error function should be of the form:

$$\sum_{f \in L} f_e(d(f), \bar{d}(f))$$

Closer inspection of the data shows that a lot of serious errors in prediction appear next to the glottal stop. Glottal stop has not been labelled and its duration is added to the duration of one of its neighbors. Labelling glottal stops would probably improve the accuracy of the model.

Other observations show that the large number of errors occur in words with high prominence. Some researchers (see e.g. [7]) report that prominence (both syllable and word prominence) can be used as an important factor influencing not only duration, but other prosodic attributes as well. Thus, prominence is another candidade for the further investigation.

8. Conclusions

A duration model for Czech text-to-speech synthesis was presented in the papers. It is based on the CART modeling technique. Several versions of the model have been presented. The best solution combines regression trees built separately for three different phoneme classes — vowels, sonorant consonants and other consonants. The correlation coefficient of the best model was 0.79 and RMSE 20.3 milliseconds. The model will be applied in the Czech speech synthesizer Demosthenes.

9. References

[1] Batůšek. R. An objective measure for assessment of the concatenative tts segment inventories. In *Proceedings of Eurospeech 2001 — Scandinavia*, Aalborg, Denmark, Sept. 2001.

[2] Batůšek, R. and Dvořák, J.. Text preprocessing for czech speech synthesis. In *Proceedings of TSD'99*, Pilsen, Czech Republic, Sept. 1999.

[3] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. *Classification and Regression Trees*. Chapman Hall, New York, USA, 1984.

[4] Chung, H. and Huckvale, M. A. Linguistic factors affecting timing in korean with application to speech synthesis. In *Proceedings of Eurospeech 2001 — Scandinavia*, Aalborg, Denmark, Sept. 2001.

[5] Lee, S. and Oh, Y.-H. Tree-based modeling of prosodic phrasing and segmental duration for korean tts system. *Speech Communication*, 28(4), 2000.

[6] Czech SAMPA www page. http://noel.feld.cvut.cz/sampa. WWW page, Nov. 2001.

[7] Portele, T. and Heuft, B. Towards a prominence-based synthesis system. *Speech Communication*, 21, 1997.

[8] Shih, C. and Ao, B. Duration study for the bell laboratories mandarin text-to-speech system. In J. P. H. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, editors, *Progress in Speech Synthesis*, chapter 31, pages 383–400. Springer, Berlin, Germany, 1996.

[9] van Santen, J. P. H. Assignment of segmental duration in text-to-speech synthesis. *Computer, Speech and Language*, 8, 1994.