TRAMES, 2007, **11**(**61**/**56**), 2, 284–298

MODELLING SPEECH TEMPORAL STRUCTURE FOR ESTONIAN TEXT-TO-SPEECH SYNTHESIS: FEATURE SELECTION

Meelis Mihkla

Institute of the Estonian Language, Tallinn

Abstract. The article discusses the principles of selecting features for modelling the temporal structure of Estonian speech, using different types of read-out texts, with a view to text-to-speech synthesis (TTS). Feature selection is known to depend on certain general issues regulating speech temporal structure, as well as on some language specific aspects. The durational model of Estonian stands out for some foot-bound features (foot quantity degree, number of feet in the word) being included in the input. In addition to the traditional descriptors of sound context and hierarchical position the prediction of Estonian segmental durations requires information on some morphological, syntactic and lexical features of the word, such as word form, part of sentence, and part of speech. In the prediction of pauses in the speech flow the relevant features are: distance from sentence beginning and from the previous pause, the length and quantity degree of the preceding foot, and the occurrence of a punctuation mark or conjunction. Although expert opinions were used in feature selection, statistical methods should be applied to test the vector of optimal argument features.

Keywords: feature selection, speech timing, segmental durations, pauses, text-to-speech synthesis, feature significance, statistical modelling

1. Introduction

Variability is one of the keywords of speech technology. While in speech recognition variability in the speech wave is a frequent source of trouble, insufficient variability in speech synthesis may lead to monotony and unnaturalness of synthetic speech (Tatham, Morton 2005:9). Synthetic speech cannot boast a natural temporal structure without successfully rendering the normal variability in the duration of sounds and pauses, as well as in the positioning of pauses in the speech flow. This is a complex task requiring an optimal choice of features to model speech timing. Modelling speech temporal structure is of particular interest for speech technology, representing an interface between the cognitive and mechanical aspects of speech generation. The general aspects governing that structure derive from the human vocal tract and articulatory mechanisms being the same for all languages. The specific aspects, however, are connected with factors typical of the language and the speaker. It is, for example, highly probable that one and the same sound sequence as pronounced by two different speakers (or even by the same speaker on two different occasions) has different timing characteristics (Campbell 2000:281).

Generation of synthetic speech with a prosodically appropriate temporal structure is never easy as speech prosody is subject to the influence of many factors, often with complex joint effects. Moreover, a language may bring forth a lot of factorial coincidence resulting in a surprising number of exceptional cases (van Santen 1998:115). Characteristics controlling speech timing have been studied for a long time and in several different speech-related spheres. However, not all discoveries have as yet been circulated in full, neither have all factors been integrated in a single extensive model to be used in all relevant spheres. Even confining oneself to just one concrete sphere, it is hard to comprehend the underlying principles and mechanisms of the characteristics (Sagisaka 2003:1).

Over the past decade the developments in speech synthesis have revealed a certain tendency of unification and multilingualism, which means that similar development systems, methods and approaches are applied to many different languages. The existing Estonian text-to-speech synthesizer, for example, uses the MBROLA synthesis engine worked out at Mons University, Belgium (Dutoit 1997:276). The question remains, however, to what extent unified algorithms and unified feature selection could be applied in speech prosody, and in particular, in the modelling of speech timing.

Characteristics of segmental durations have been measured for many languages in order to pinpoint the universal and the specific in the timing patterns of languages (Sagisaka 2003:1). Speech technology has been striving to obtain precise control over segmental durations in order to synthesize speech with a natural-sounding rhythm and timing. The present paper is focused on the selection of optimal characteristics to model speech temporal structure for Estonian text-to-speech synthesis.

2. Principles of feature selection for modelling timing in speech

There have been three approaches in the control of speech timing: one is mora-timed rhythm, which is used, e.g., in Japanese, the second is syllable-timed rhythm – every pronounced syllable is supposed to take up roughly the same amount of time, and the third is stress-timed rhythm, recognized and used in many Germanic languages.

In Japanese, mora isochrony has been observed as a temporal constraint controlling vowel duration. A negative correlation has been found to exist between the durations of vowels and their neighbouring consonants. The fact that the temporal compensation of the duration of a vowel is more influenced by the duration of its preceding consonant is regarded as an acoustic manifestation of mora-timing. As has been proved by statistical analysis, such compensation takes place in mora units, not in syllables (Sagisaka 2003:2). This does not of course exclude more extensive regulation. Speech rhythm is readjusted on phrase level - the more moras in the phrase, the shorter their average durations. In the end it is the mora constraints and the local adjustment of speech rate within phrases that determine most of the variation of segmental durations in read-out Japanese speech (Sagisaka 2003:2). Mora metrics has been applied quite successfully in Estonian phonology as well. In Estonian word prosody Arvo Eek has interpreted intra-foot quantity as a manifestation of mora isochrony, where the quantity degree is determined by the distribution of durations within the foot (Eek, Meister 2004:336–357).

In a syllable-timed language, every syllable is thought to take up roughly the same amount of time when pronounced, though the actual duration of a syllable depends on the situation. Spanish and French are commonly quoted as examples of syllable-timed languages. When a speaker repeats the same sentence many times at the same rate of articulation, the durations of adjacent phones display a strong negative correlation, i.e. any variance in the duration of a single phone is compensated in the adjacent phones and so the temporal sequence of articulation must be organised at levels higher than phoneme, e.g. syllable (Huggins 1968). The syllable-timing hypothesis was proposed by Campbell and Isard in statistical modelling to account for the interaction between higher and lower levels of timing control (Campbell and Isard 1991). It posits the syllable as a mediator and offers a way to map the effects of linguistic and semantic contexts onto the physiological constraints of speech sound production. By adopting a higher level framework for duration control, it overcomes the sparcity of data problem in the modelling of the variability in individual phone durations (Campbell 2000:307).

In a stress-timed language, syllables may last different amounts of time, but there is a constant amount of time (on average) between two consecutive stressed syllables. For English the rules of speech timing were formulated by Dennis Klatt. Using the results of other researchers he made up 11 rules describing 84% of segment temporal variation in a text read out by himself (Klatt 1979). Klatt's rule-based model has been modified and developed by other researchers. Jan van Santen generalized the Klatt rules, adding the following six factors affecting segmental durations in American English (van Santen 1998:123–124):

1. phonetic segment identity (30 values),
2. identities of surrounding segments (10),
3. syllabic stress (3),
4. word 'importance' (contrastive stress),
5. location of the syllable in the word,
6. location of the syllable in the phrase

Thus the above models are centred on the stress groups (syllable stress, contrastive stress). Rule-based models were good enough to provide reasonable segment durations for most cases, yet sometimes grave mistakes occurred. The mistakes were often due to attempts of simultaneous application of independently derived rules. The advent of large databases, however, enabled the use of statistical methods to predict segmental durations much more precisely. Many statistical models have made sound use of the parameters and features of Klatt's rule-based model. Those have been applied as argument features, either directly or selectively, integrating some language-specific information. Horák, for example, introduced a special feature of monosyllabic words in his model of Czech durations (Horak 2005:79), and Vainio roped in some morphological features and part-of-speech information while modelling prosody for Finnish TTS (Vainio 2001:66–67). Timing control is just an aspect depending on language-specific features.

3. Feature selection for Estonian TTS

Chapter 2 described three basically different phonological approaches to feature selection for a durational model: one was based on mora metrics, the second on syllables and the third on stress-timing. Although certain characteristics allow for Estonian being included among mora-counting languages (Eek, Meister 2004:336), our system of features will be based on stress and quantity degree, considering the main characteristics of syllable and foot structure in Estonian language.

The central unit of Estonian prosody is the foot, carrying three quantity degrees (Q1, Q2, Q3) as the phonologically relevant prosodic oppositions. The quantity degree is a suprasegmental feature resulting from the joint effect of several other features, one of the most important of which is the ratio of the durations of syllables or their components[1]. Estonian quantity degrees and stress are usually described in the framework of a prosodic hierarchy enabling to divide an utterance into components lying on different levels of subordination (Eek, Meister 2004:253). As can be seen in Fig. 1, a sentence or phrase consists of prosodic words, while the words, in turn, consist of feet, the feet consist of syllables and the rear is brought up by phonemes, making up the lowest segmental level. As in Estonian sentences, phrases (noun phrase, verb phrase, adverbial phrase) are often quite closely intertwined, we define a phrase for the present paper as a finite clause or a list element bounded by an intra-sentence punctuation mark or conjunction. Thus, Fig. 1 may apply equally to either a sentence or a phrase.

The relative position of a current phone is rendered in a hierarchic scale as follows: position of the phone in the syllable, position of the syllable in the foot, position of the foot in the word, position of the word in the phrase. In addition, as has been proved by previous analysis, information on sentence and word length is necessary.

The next underlying principle of feature selection states that every phone has its intrinsic duration, while at the same time the phone is affected by its neighbouring phones. Intrinsic durations and phone interaction have been studied for many languages. Such universal linguistic phenomena are also observed in Estonian. The first measurements of the intrinsic durations of Estonian vowels took place about half

[1] The quantity degree is defined as a ratio of structural components of stressed and unstressed syllables in the form σ_{stressed}(nucleus+[coda]) / $\sigma_{\text{unstressed}}$(nucleus).

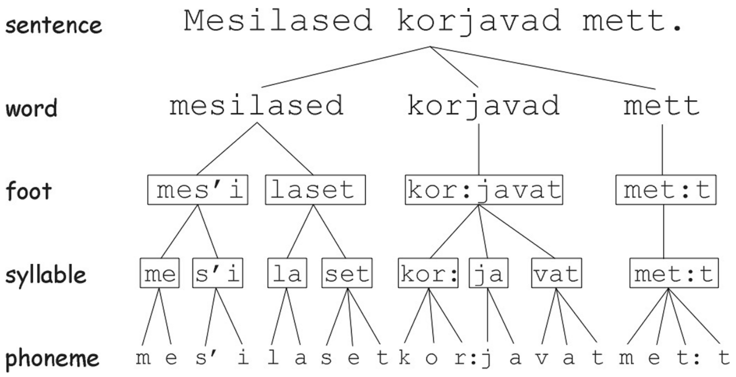

Figure 1. 'Bees gather honey' – hierarchical encoding of the relative position and length of a current speech unit. For example the phone [l] is encoded according to its position in the syllable [la] of a length of two phones. The position of the syllable [la] is encoded in relation to the foot [laset] with a length of two syllables. The foot is further assigned a code according to its place in the word [mesilased] etc.

a century ago (Liiv 1961). In several later studies of intra-phone microprosodic variations it has been stated that in Estonian the short low vowels are about 10–15 ms longer than the high vowels (Eek, Meister 2003:836; Meister, Werner 2006:111). Interaction between neighbouring phones is manifested in consonant shortening in clusters, in particular in the neighbourhood of voiceless consonants (Eek, Meister 2004:267).

The current phoneme segment is characterized by phoneme identity and phoneme length. In Estonian there are 9 vowel phonemes and 17 consonant phonemes (Eek, Meister 1999). The class and length of the left and right neighbours are also important. In modelling, the basic question is how many left and right neighbours affect the duration of a current phone. Usually their number is 1, 2 or 3. Experimental modelling of Estonian phone durations has shown that it takes two phonemes from the left (previous and previous but one) and two from the right (next and next but one) to achieve an optimal description of the context of a current phoneme (see Fig. 2). The phonemes are defined by their phoneme class (9 classes, pause included) and contrastive length (short vs. long). A phoneme and its context takes 10 features to describe, while the hierarchical position of the phoneme in the utterance is encoded by 5 features, some speech units (syllable stress, syllable type, foot quantity degree) take 3 features, and the length of higher-level units (syllable, foot, word, phrase, sentence) needs 5 features. There is also a binary feature referring to a punctuation mark. All these features (24 in all) make up a vector of basic features to serve as input for the durational model. Another point to consider when selecting initial features was their being supported by the technologies available for the Estonian language, enabling the features to be generated automatically from the input text. For the present study we have made use of a sentence builder, syllabifier, morphological analyzer, disambiguator a.o. modules provided by Estonian language technologists (Viks 2000), (Kaalep, Vaino 2001).

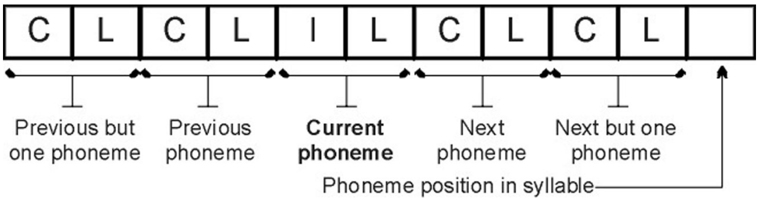

Figure 2. Encoding information on current phoneme neighbourhood (C – phoneme class, L – phoneme contrastive length, I – phoneme identity).

3.1. Initial data and the modelling environment

As our aim was to model speech temporal structure for a text-to-speech synthesizer, the analysis was based on read-out texts. A one-to-one correspondence between text and speech enables transition from a symbolic representation of prosody to an acoustic one as well as to find out to what extent, if at all, the syntactic structure of the written text could be related to the prosodic structure of speech.

The training material consisted of speech passages from a CD-version of a mystery story (Stout 2003) read by a professional actor, speech passages from longer news texts read by announcers of the Estonian Radio, and speech passages from the BABEL Estonian phonetic database (Eek, Meister 1999). In total there were over 60 speech samples read by 27 speakers , while the samples lasted from half to two minutes. All those samples were manually segmented into phones and pauses.

The speech temporal structure was modelled statistically using the Enterprise Miner workspace of SAS 9.1 software (see Fig. 3 for a block diagram of the data processing).

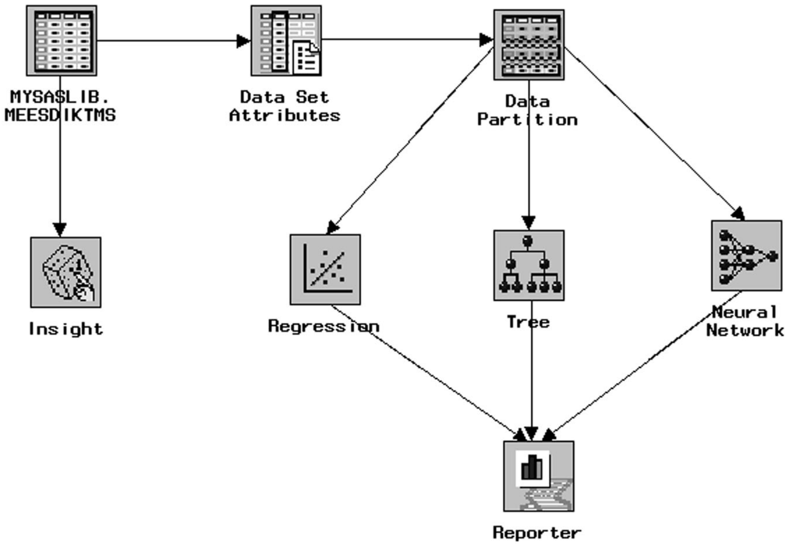

Figure 3. SAS Enterprise Miner workspace.

Linear regression, CART and neural networks were used as methods of prediction. In modelling, the material of each speaker's total presentation was divided into three parts: 50% was assigned for training the model, while 30% was used for data validation and the remaining 20% was meant for testing.

3.2. Expert opinions

After a rough selection has been made it would be quite useful to have a few experts give their opinions on the resulting vector of argument features as well as a few recommendations on possible additional features. The experts could estimate whether or not a feature is significant in the prediction of speech timing (e.g. segmental durations), also, it would be useful to have their opinion on the possible joint effects of certain feature constellations. During our first experiments in statistical modelling we invited Estonian phoneticians and speech technologists to evaluate our first vector of argument features. The coincidence between their opinions and our preliminary results was a mere 41–65% (Mihkla, Kuusik 2005:95). Due to the interim accumulation of speech material, however, and the increasing volume of Estonian prosodic speech corpora our recent results are in better harmony with expert opinions (see Table 1). Part of the – still considerable – difference may be due to the fact that most of the 'duration patterns' of the phoneticians are based on measurements of isolated words and sentences, whereas our results draw on fluent speech. Sound

Table 1. Expert opinions versus results of regression analysis (ExpN- N expert, Reg – results of regression analysis, 1 – significant explanatory variable, 0 – insignificant variable)

Explanatory variable	Exp1	Exp2	Exp3	Exp4	Reg
Previous phoneme class	0	0	0	0	1
Previous phoneme length	1	1	1	1	1
Current phoneme identity	1	1	1	0	1
Current phoneme length	1	1	1	1	1
Next phoneme class	1	1	0	0	1
Next phoneme length	1	0	1	1	0
Phoneme position in syllable	1	1	0	1	1
Stress of syllable	1	1	1	1	1
Type of syllable	1	0	0	1	0
Quantity degree of foot	1	1	1	1	1
Syllable position in foot	1	1	1	1	1
Length of foot in syllables	1	1	0	1	1
Foot position in word	1	0	0	1	0
Length of word in feet	1	1	0	1	1
Word position in phrase	1	1	1	1	1
Length of phrase in words	1	0	0	1	1
Length of sentence in phrases	1	0	0	0	1
Total of 'correct' answers	13	14	9	10	
%	76%	82%	53%	59%	
Total average %				67%	

durations measured on isolated sentences and the temporal structure of fluent speech, however, are known to differ quite considerably (Campbell 2000:312–315).

3.3. Lexical prosody

Traditionally a list of factors significantly affecting speech timing does not include either part-of-speech (POS) information or morphological characteristics (van Santen 1998, Campbell 200, Sagisaka 2003). This may be due to most studies on TTS synthesis focusing on languages with relatively little morphology. Finnish is one of the few languages boasting a study of the influence of morphological features on the duration of speech units (Vainio 2001). In Estonian the word has a very important role both in grammar and phonetics, while the morphology is extremely rich. Hence our interest to check whether there are any morphological, lexical, or even syntactic features possibly affecting the temporal structure of Estonian speech. The most natural way to find out that information seemed to lie through an extension of our earlier methodology of statistical modelling to see how certain morphological, lexical and syntactic characteristics might affect the functioning of our durational models. The modelling was done using two different methods - linear regression and a nonlinear method of neural networks. Change of the output error was measured to enable qualitative assessment of the influence of the factors under study. The results demonstrated a couple of percent error decrease in case some morpho-syntactic and POS information had been added to model input (Mihkla 2007).

As the models were based on the speech of merely two radio announcers it seemed a little premature to generalize the possible interpretations. It should however be mentioned that the most distinct regularities were revealed by a visual observation of the POS regression coefficients. Table 2 demonstrates the mean lengthenings and shortenings, by part of speech, of speech sounds relative to verb sounds in the durational models of male and female speech. As we can see, there is more variation in the middle part of the table, while the top and bottom parts are rather similar. Table 2 reveals that in proper names sounds are pronounced longer by 5.2–6.2 ms on average. The two newscasters' mean phone length was 62.5 and 64.1 ms respectively. Consequently their pronunciation of proper names was about 10% longer than verbs. Of the latter, nouns and adpositions were pronounced a little longer. It was surprising to find such lengthening in adpositions as in most languages function words are shorter than content words. An Estonian adposition invariably belongs to a noun phrase. The noun often stands in the focus of the sentence, while its more than average length may extend to a neighbouring adposition. Ordinal numerals, however, were pronounced over 10% shorter, and pronouns and adverbials ca. 5% shorter than average. The shortening of ordinal numerals can be accounted for by quite many dates in the text, which are typically expressed by ordinal numerals. Reading the relatively long dates of the past century the newscasters tend to hurry. This is because usually only the last one or two numbers of the year are important, but nevertheless the whole number has to be pronounced, as required by rules of correct reading.

Table 2. The average lengthening-shortening values (in ms) of sound durations for different parts of speech in the male and female material

Part of speech	Male announcer	Female announcer
Proper noun	6.23	5.22
Noun	2.25	2.10
Adposition	0.82	2.82
Genitive attribute	0.42	1.35
Verb	0.00	0.00
Numeral	–0.10	0.42
Conjunction	–0.14	1.81
Adjective	–0.39	1.14
Adverb	–0.89	–2.90
Pronoun	–4.13	–3.86
Ordinal numeral	–5.44	–7.48

3.4.Pauses in speech

Regulating the primary syntactic division or prosodic phrasing with an aim to facilitate comprehension of the utterance, pauses are one of the factors determining speech rhythm (Tseng 2002). Despite the high variability of pauses in natural speech, different kinds of pauses vary in duration. At least in texts read aloud at a normal speech rate it is possible to distinguish between phrase-, sentence- and paragraph-final pauses by their duration, as has been proved by statistical analysis (Mihkla 2006a:290). A natural-sounding rhythm of synthetic speech would mean a good enough rendering of both the duration of pauses and their location in the speech flow.

For modelling pause durations some texts were used to generate a number of characteristics describing the following:

- text structure (end of paragraph, sentence, or phrase; conjunctions within the text)
- prepausal foot (length of the foot in phones, foot quantity degree, length of the foot-final syllable in phones and a binary characteristic indicating final lengthening)
- pause timing specifications (distance of the pause from the beginning of the paragraph, sentence, and phrase, as well as its distance from the previous pause and the previous breathing).

The feature to be predicted was pause duration. The use of linear regression required the response to be logarithmed, for logarithmed values are more likely to yield a normal distribution. See Fig. 4 for a regression tree to calculate pause duration.

The first-level classification of pauses on the regression tree distinguishes between sentence-final and non-sentence-final pauses. The intra-sentence pauses branch off to the left, while sentence-final ones go to the right. The intra-sentence pauses are, in turn, dichotomized depending on whether they happen to finish a phrase or not. The length of an intra-phrasal pause, for example, distanced from the previous pause by less than seven feet is 166 ms, whereas a longer distance correlates with a 253 ms pause.

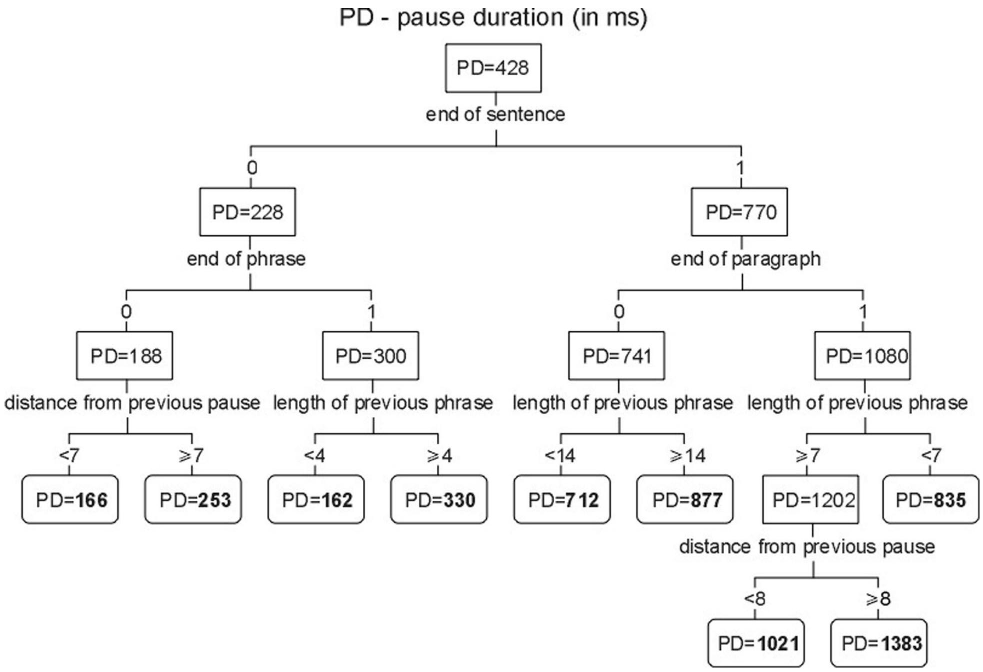

Figure 4. Regression tree to model pause durations in a text. The boxes with rounded corners stand for the leaves of the tree, each referring to the duration of the appropriate pause class.

Prediction of pause location was started by applying logistic regression, meant to predict the probability of a pause following a given word in the speech flow. The input variables were the same as used for predicting pause durations. There were two additional binary features, indicating whether the following word is a proper noun or a foreign word. Their addition was inspired by the idea that there might be a short pause before the pronunciation of proper names (e.g. *Minu nimi on Tamm, Jüri Tamm* 'My name is Tamm, Jüri Tamm') and maybe also before some more sophisticated foreign words (e.g. *Rahvas toetas konstitutsioonilist monarhiat* 'The people supported constitutional monarchy'). That hypothesis was, however, disproved. The correlation of pauses with proper names and foreign words was extremely weak and thus the features proved insignificant.

Table 3. Pause locations as predicted by the logistic model

	Correct predictions	Actual no. of pauses in the sample	Percent of correctness
Pause after word (PAUSE = 1)	402	600	67
No pause (PAUSE=0)	2510	2708	93
TOTAL			88

As can be seen from Table 3 the model predicted correctly the location of 67% of the pauses. The total predictive precision of the model, i.e. its ability to guess whether a certain word will be followed by a pause or not, amounted to 88%. Analysis showed that certain markers of text structure (end of paragraph, end of sentence, colon, dash) are followed by a pause with a 93-100% probability. Leaving out such well-marked pauses the model was left with a mere 44% of predictive power.

3.5. Feature significance, predictive precision and errors

Prior to evaluating the significance of features one should test the statistical relevance of the model. Table 4 shows an example of the summary of fit and variability analysis of our regression model of pause durations. We can see that the model is statistically significant and it describes ca 67% of the variability of pause durations (R-square = 0.6686). Other durational models described 65–73% of pause duration variability, while for phone durations the reading amounted to 52–63%.

Tabel 4. Summary of fit and variability analysis of a regression model of pause durations

<table>
<tr><td colspan="6">Summary of fit</td></tr>
<tr><td colspan="3">Mean of response –0,86673</td><td colspan="3">R-square 0.6686</td></tr>
<tr><td colspan="6">Analysis of variance</td></tr>
<tr><td>Source</td><td>DF</td><td>Sum of squares</td><td>Mean square</td><td>F stat</td><td>Pr > F</td></tr>
<tr><td>Model</td><td>9</td><td>478.5</td><td>18.403</td><td>220.87</td><td><0.0001</td></tr>
<tr><td>Error</td><td>560</td><td>408.8</td><td>0.0862</td><td></td><td></td></tr>
<tr><td>C Total</td><td>559</td><td>787.2</td><td></td><td></td><td></td></tr>
</table>

Feature significance is best determined in the case of a regression model, where the significance of each feature gets a statistical evaluation. The method of forward selection, for example, means that the most significant features at the moment are added to the model one by one, with revaluation taking place before each cycle. CART also makes sure that the regression tree will get the most significant features. In neural networks, however, there is no such evaluation of significance and the relevance of a feature can be estimated manually, by adding or removing features one by one and evaluating the output for each case.

Table 5 contains the features that – on the basis of extensive experimental material – have been found to be significant in predicting phone duration. The number of significant features may vary across speakers; it also depends on the used method. The features in bold letters on dark grey background mark features that were significant for all speakers. The features in normal lettering on a lighter grey background were insignificant for the durational models of some speakers. Surprisingly the latter group includes the quantity degree of foot, which is considered a cornerstone of Estonian speech prosody. One of the possible reasons might lie in the circumstance that quantity degree as a suprasegmental feature cannot be represented as a single linear characteristic, but rather as a ratio of structural components of stressed and unstressed syllables in the form σ_{stressed}(nucleus+[coda]) / $\sigma_{\text{unstressed}}$(nucleus). Another fact suggestive of possible mutual effects between stress and syllable structure is that syllable stress does not always correlate significantly with the duration to be predicted. Surprisingly enough the contrastive length of the next phone happened to be less significant than that of the next but one. The normal type on a white background marks the two features that proved systematically insignificant for the prediction of segmental durations, notably, type of syllable (open or closed) and the position of the foot in the word.

Table 5. Significance of input features for modelling segmental durations

Inputs (per *sound)*	
1.	**previous but one phoneme class**
2.	**previous but one phoneme length**
3.	**previous phoneme class**
4.	previous phoneme length
5.	**current phoneme identity**
6.	**current phoneme length**
7.	**next phoneme class**
8.	next phoneme length
9.	**next but one phoneme class**
10.	**next but one phoneme length**
11.	**phoneme position in syllable**
12.	stress of syllable
13.	type of syllable
14.	quantity degree of foot
15.	**syllable position in foot**
16.	**length of foot in syllables**
17.	foot position in word
18.	**length of word in feet**
19.	**monosyllabic word**
20.	**word position in phrase**
21.	**length of phrase in words**
22.	length of sentence in phrases
23.	**punctuation**
24.	**morphology**
25.	**part-of-speech**
26.	**part of sentence**

Table 6 contains the six features proved by logistic regression to affect the positioning of an intra-sentence pause. A comma in the text is very important, raising the chances for a pause to occur in speech by 17.4 times. A 7–8 times higher chance for the word to be followed by a pause is signalled by a following conjunction or a lengthened final foot. Slightly more frequently than average a pause can be expected to occur after longer feet or after words of longer quantity degrees. In a predictive model, however, the role of the latter two features remains relatively marginal, raising the chances for a pause to occur by no more than 1.2–1.3 times.

Table 6. Results of logistic regression: variables explanatory of the location of an intra-sentence pause, their ratio of chances and confidence levels

Independent variables	Odds ratio	Confidence levels	
		Lower	Upper
The word is followed by a comma	17.4	11.7	25.9
The next word is a conjunction	7.9	4.8	12.8
Distance of the word from sentence beginning	1.1	1.0	1.2
Length of the preceding foot	1.3	1.1	1.4
Quantity degree of the preceding foot	1.2	1.1	1.5
Lengthening of the preceding foot	6.9	5.2	9.2

Depending on the speaker and the method, the predictive error of phone durations was within 16.1–21.2%. Testing different methods (linear regression, CART, neural networks) on one and the same data set, it turned out that the predictive precision of linear regression and the neural networks was nearly equal, while the CART model had a slightly higher predictive error than the rest (Mihkla 2006b:123). We were surprised to find that the linear method could compete with non-linear ones, although a linear model is usually expected to show nothing but the most obvious and most general relations between input and output, and only nonlinear methods are trusted to reveal the more covert relations.

For pause durations the predictive error was - depending on the speaker - 8-12%. The predictive precision of the logistic model for pause locations is presented in Table 3. If, however, the punctuation-bound pauses are left out, the predictive precision of intra-sentence pauses drops to 44%.

4. Conclusion

Generation of synthetic speech with a prosodically appropriate temporal structure is complicated as speech prosody is influenced by many factors. In feature selection one should consider certain general aspects governing speech timing as well as some language specific ones. The Estonian durational model stands out for certain foot-bound features (foot quantity degree, number of feet in the word) being included in the model input. Although feature selection involved the use of expert opinions, too, the vector of optimal argument features should definitely be tested by statistical methods. The considerable difference between the expert opinions and the results of fluent speech analysis may be due to the fact that the ‘durational patterns’ underlying the decisions of the expert phoneticians were quite likely based on measurements of laboratory speech (i.e. isolated sentences and words). Prediction of pause durations and locations in the speech flow was proved to heavily depend on the following features: distance of the word from sentence beginning and the previous pause, length of the preceding phrase, length and quantity degree of the preceding foot, and the occurrence of a punctuation mark or conjunction. For the durations of speech units the predictive precision of the model differed across the methods used as well as speakers, while for segmental durations the predictive error was 16–21% and for pause durations it was 8–12%. Although in the logistic model the summary estimation of the possible occurrence of a pause after a text word amounted to 88%, predictive precision for intra-sentence pauses did not exceed 44%.

Apart from the traditional parameters, describing the context of a phone and its hierarchic position in the sentence, prediction of Estonian segmental durations requires an addition of certain morphological, syntactic and lexical features such as word form, part of sentence and part of speech. The most distinct regularities were revealed by part-of-speech analysis of the durational model, showing that proper names and nouns were pronounced the longest, whereas the least time was spent on ordinal numerals and pronouns.

Acknowledgements

The study was financed by the national programme „Language technological support of Estonian”. My heartfelt thanks to Arvo Eek (Institute of Cybernetics at TTU, Tallinn) and Einar Meister (Institute of Cybernetics at TTU, Tallinn) for their valuable remarks and comments.

Address:
Meelis Mihkla
Institute of the Estonian Language
Roosikrantsi 6
10119 Tallinn, Estonia
Tel.: +372 6446 947
E-mail: meelis@eki.ee

References

Campbell, Nick (2000) “Timing in speech: a multilevel process”. In *Prosody: theory and experiment,* 281–334. M. Horne, ed. Dordrecht/Boston/London: Kluwer Academic Publishers.

Campbell, N. W. and S. D. Isard (1991) “Segment durations in a syllable frame” *Journal of Phonetics* 19, 37–47.

Eek, Arvo and Einar Meister (1999) “Estonian speech in the BABEL multi-language database: phonetic-phonological problems revealed in the text corpus”. In *Proceedings of LP’98,* II, 529–546. O. Fujimura, ed. Prague: The Karolinum Press.

Eek, Arvo and Einar Meister (2003) “Foneetilisi katseid ja arutlusi kvantiteedi alalt (I): Häälikukestusi muutvad kontekstid ja välde”. [Phonetic tests and disputes about quantity (I): Contexts changing sound duration and quantity degree.] *Keel ja Kirjandus* (Tallinn) 46, 11, 815–837 and 12, 904–918.

Eek, Arvo and Einar Meister (2004) “Foneetilisi katseid ja arutlusi kvantiteedi alalt (II): Takt, silp ja välde”. [Phonetic tests and disputes about quantity (II). Foot, syllable and quantity.] *Keel ja Kirjandus* (Tallinn) 47, 4, 251–277 and 5, 336–357.

Dutoit, Thierry (1997) *An introduction to text-to-speech synthesis*. Dordrecht: Kluwer Academic Publishers.

Horak, Pavel (2005) “Using neural networks to model Czech text-to-speech synthesis”. In *Proceedings of the 16th Conference of electronic speech signal processing,* 76–83. R. Vich, ed. Prague: TUDpress.

Huggins, A.W.F. (1968) “The perception of timing in natural speech: compensation within syllable”. *Language and Speech* 11, 1–11.

Kaalep, Heiki-Jaan and Tarmo Vaino (2001) “Complete morphological analysis in the linguist’s toolbox”. In *Congressus Nonus Internationalis Fenno-Ugristarum, Tartu 7.-13.08.2000,* V, 9–16. Tartu: TÜ Kirjastus.

Klatt, D. H. (1979) “Synthesis by rule of segmental durations in English sentences”. In *Frontiers of Speech Communication research,* 287–300. B. Lindblom and S. Öhman, eds. New York: Academic Press.

Liiv, Georg (1961) “Eesti keele kolme vältusastme vokaalide kestus ja meloodiatüübid”. [Duration of vowels of the three quantity degree of Estonian and types of melody.] *Keel ja Kirjandus* (Tallinn) 4, 7, 412–424 and 8, 480–490.

Meister, Einar and Stefan Werner (2006) “Intrinsic microprosodic variations in Estonian and Finnish: acoustic analysis”. In *Fonetiikan Päivät 2006 = The Phonetics Symposium 2006,* 103–112. R. Aulanko, L. Wahlberg, and M. Vainio, eds. (Publications of the Department of Speech Sciences, University of Helsinki) Helsinki: University of Helsinki.

Mihkla, Meelis and Jüri Kuusik (2005) “Analysis and modelling of temporal characteristics of speech for Estonian text-to-speech synthesis”. *Linguistica Uralica* 41, 2, 91–97.

Mihkla, Meelis (2006a) “Pausid kõnes”. [Pauses in Speech.] *Keel ja Kirjandus* (Tallinn) 49, 4, 286–295.

Mihkla, Meelis (2006b) “Comparison of statistical methods used to predict segmental durations”. In Fonetiikan Päivät 2006 = The Phonetics Symposium 2006, 120–124. R. Aulanko, L. Wahlberg, and M. Vainio, eds. (Publications of the Department of Speech Sciences, University of Helsinki) Helsinki: University of Helsinki.

Mihkla, Meelis (2007) “Morphological and synthetic factors in predicting segmental durations for Estonian text-to-speech synthesis”. *Proceedings ICPhS 2007.* (accepted, in print).

Sagisaka, Yoshinori (2003) “Modeling and perception of temporal characteristics in speech”. In *Proceedings of 15th International Congress of Phonetic Sciences,* 1–6. M. J. Sole, D. Recasens, and J. Romero, eds. Barcelona.

van Santen, Jan (1998) “Timing”. In *Multilingual text-to-speech synthesis: the Bell Labs approach*, 115–140. R. Sproat, ed. Kluwer Academic Publishers.

Stout, Rex 2003 “Deemoni surm”. [Death of a Demon.] CD-versioon (Read by Andres Ots). Tallinn: Elmatar.

Tatham, Mark and Katherine Morton (2005) *Developments in speech synthesis*. Chichester: John Wiley & Sons Ltd.

Tseng, C. (2002) “The prosodic status of breaks in running speech: examination and evaluation”. In *Proceedings of Speech Prosody 2002,* 667–670. Aix-en-Provence, France.

Vainio, Martti (2001) *Artificial neural network based prosody models for Finnish text-to-speech synthesis*. Helsinki: University of Helsinki.

Viks, Ülle (2000). “Eesti keele avatud morfoloogiamudel” [Open morphology model of Estonian language.]. In *Arvutuslingvistikalt inimesele*, 9–36. [From computational linguistics to people.] T. Hennoste, ed. (Tartu Ülikooli üldkeeleteaduse õppetooli toimetised, 1.) Tartu.