Linguistica Uralica XLI 2005 2

MEELIS MIHKLA, *JÜRI KUUSIK* (Tallinn)

ANALYSIS AND MODELLING OF TEMPORAL CHARACTERISTICS OF SPEECH FOR ESTONIAN TEXT-TO-SPEECH SYNTHESIS*

Abstract. A text-to-speech system must be capable of generating sounds and pauses with such durations that do not noticeably differ from natural speech. Currently, the prosodic modelling of Estonian text-to-speech synthesis is largely based on generalized measurements of speech units in isolated words and sentences, and as a result the synthesized speech is often monotonous and has poor fluency. In this work the first attempts are made to improve the naturalness of the output speech of the speech synthesiser with the help of statistical duration models of fluent speech. The source material consisted of (a) prose read out by a professional actor, and (b) news broadcasts read by announcers. On the basis of this material variability of the duration of pauses and boundary lengthenings was investigated. It turns out that in the case of a read text at normal speech rate the classification of speech pauses is perfectly possible and can be applied in speech synthesis. An attempt was also made to establish whether and to what extent the syntactic parsing of a text is related to the prosodic parsing of speech. A generalized regression analysis revealed what features are essential in predicting sound durations in speech and a statistically optimal model was developed. Curiously the quantity degree of a foot, despite being the cornerstone of Estonian word prosody, was not a significant feature for prediciting the duration of a sound on the basis of this material. The results of the modelling were then compared with the expert opinions of some Estonian phoneticians.

Keywords: Estonian, duration of sounds, pause, statistical modelling, regression analysis, text-to-speech synthesis.

1. Introduction

The task of text-to-speech synthesis is to convert orthographic text into natural-sounding speech. For the artificial speech to sound realistic to the human ear, it should comprise realistic intonation, rhythm and stress patterns. More specifically, the text-to-speech system must be able to generate durations of sounds and pauses not notably different from the values of the actual speech.

Currently, prosodic modelling in Estonian text-to-speech synthesis (Mihkla, Meister, Eek 2000) is largely based on generalized measurements

* Support from the Estonian Science Foundation, grant No. 5039, and state program "Estonian language and national memory" has made the present work possible.

of speech units in isolated words and sentences. The resulting output (synthesized) speech, however, is often monotonous and has poor fluency, which sets application limitations on the synthesizer. As indicated by Nick Campbell, the durations of sounds in isolated words or sentences are largely different from durations of sound in the fluent speech (Campbell 2000). The speech contains complicated temporal patterns, which the text-to-speech system must be able to imitate for the speech to sound natural. The availability of oral speech corpuses provides an opportunity to achieve the text-prosody transformation with the help of statistical models.

In this work the first attempts are made to improve the naturalness of the output speech of an Estonian speech synthesiser with the help of statistical duration models of fluent speech. We applied the technology of regression analysis to find out the essential features of sound durations and to compose a prediction model. The results of modelling the durations are compared with expert opinions given by Estonian phoneticians. With the aim to providing for a natural rhythm of the output speech, the relation of pauses and boundary lengthenings with syntactic parsing of the text is studied.

2. Source material

Because we are concerned with a text-to-speech synthesiser, the source material was a sample of texts read by announcers. On the basis of one-to-one correspondence of text and speech, it is possible to move from a symbol-based representation of prosody to the acoustic one and also to establish whether and to what extent the syntactic parsing of the text is related to the prosodic parsing of the speech.

The source material consisted of passages of speech from the CD-version of a detective story read by an actor (Stout 2003) and passages of speech and texts from longer news read by announcers of Estonian Radio. Altogether, 12 speech passages were analysed, each 1-2 minutes long. All passages of speech were segmented into sounds and pauses.

3. Analysis of pauses and boundary lengthenings

Prior to the application of a general statistical model, pauses and prepausal lengthenings in speech were analysed, based on this material. The pauses and prepausal lengthenings in Estonian speech have been studied cursorily or intermittently, as a by-product in the context of other tasks. Ilse Lehiste (1981) verified whether prepausal lengthenings were in correlation with the length of subsequent pauses and she established an extremely weak link. Diana Krull (1997) studied prepausal lengthenings in dialogue in two-syllable words in the context of quantity degree. Arvo Eek and Einar Meister (2003) looked at end-of-sentence lengthenings on the basis of tempocorpus. However, they examined only words of a specific structure, and focused on quantity degree features. Therefore the need became evident to measure, for Estonian language text-to-speech synthesis, pauses and boundary foot lengthenings, on the basis of a text read out from real speech.

With a view to analysing the pauses and foot lengthenings, the durations of pauses derived from the speech wave were measured, and the foot lengthenings were calculated. For the calculation of foot lengthenings, the durations of sounds comprising the foot were summed, after which the actual duration was compared to the mean duration of the given foot structure in the speech of that announcer. The first hypothesis was to verify whether pauses and foot lengthenings could be classified (for instance, whether the pauses between phrases[1] differ significantly from the sentence end or paragraph end pauses).

Table 1

Durations of pauses and boundary lengthenings (ms) in speech

Dictors	Phrase end pauses	Sentence end pauses	Paragraph end pauses	Phrase end lengthenings	Sentence end lengthenings	Paragraph end lengthenings
Actor1 (m)	352	558	1025	200	220	315
Announcer1 (f)	303	828	902	124	112	117
Announcer2 (m)	286	769	1132	95	90	122
Generalised mean	**323**	**678**	**1021**	**155**	**161**	**217**

Table 1 presents the mean durations of pauses as per announcers and the generalised mean. Looking at the generalised means suggests that in case of a text read out at normal speech rate the classification of speech pauses is fully possible. The statistical analysis of samples corroborates this surmise. Analysis of pairs of the logarithmic durations of pauses with the help of a Student t-test reveals that the t-statistic values on significance level $p = 0.01$ noticeably exceed the t-critical two-tail quantile (cf. Table 2) on probability of significance of hypothesis $p < 0.0001$. Hence it seems proved that the mean values of durations of pauses differ and the classification of pauses is fully possible, which fact could be applied in speech synthesis. The dispersion and variance however are large; therefore in speech recognition, for instance, such classification is to no avail.

When analysing, with the help of Student t-test the data of foot lengthenings (cf. Table 2) we had to accept the null hypothesis: the foot lengthenings are from samples of the same mean value.

Table 2

Student t-test results for comparison of pairs of sample means (Ph-Se — between phrase and sentence, Ph-Pa — between phrase and paragraph, Se-Pa — between sentence and paragraph)

	Pauses			Foot lengthenings		
	Ph-Se	**Ph-Pa**	**Se-Pa**	**Ph-Se**	**Ph-Pa**	**Se-Pa**
T stat	8.87	12.25	5.91	0.81	0.26	0.65
T critical two-tail	2.62	2.76	2.72	2.65	2.84	2.90
P (T <= t)	< 0.0001	< 0.0001	< 0.0001	0.42	0.79	0.52

Next examined was whether and to what extent the prosodic parsing of speech correlates with syntactic parsing where the latter is indicated by punctuation marks and conjunctions. As shown in Table 3, there is invariably a pause in speech[2] at the paragraph end and the sentence end. In case of a colon and dash too there is a strong correlation between syntax

[1] In this work, the phrase means the clause or element of enumeration, which has been determinated within the sentence by punctuation mark or conjunction.

[2] In this work we have treated as a prosodic pause an interruption of speech over 50 ms.

and prosody. Half the commas are related to pauses. The least marked in speech are phrases starting with those co-ordinating conjunctions which do not require the comma.

Table 3

Connection of pauses and foot lengthenings with the text parsing

	No. of parsings in the text	**No. of corresponding pauses in the speech**		**No. of corresponding foot lengthenings in the speech**	
		Cnt	**%**	**Cnt**	**%**
Paragraph end	21	21	100	15	71
Sentence end	58	58	100	39	67
Comma	80	41	51	42	53
Conjunction	22	7	32	12	55
Colon	7	7	100	4	57
Dash	14	13	93	13	93

Among the punctuation marks, lengthening is obviously related to the dash. Apparently the connotation is suggested by the shape of the sign — the stretched line prompts the drawl. Suggestive of the link between pauses and boundary lengthenings is the English term 'prepausal lengthening'. This term applies, on the basis of this Estonian language speech material, only 70% of the time (only 143 pauses are preceded by word or foot lengthening). According to perception tests carried out by I. Lehiste (Lehiste, Fox 1993) Estonian speakers expect significantly less final lengthening on the last syllable of the sentence than English speakers do.

But if we wish to lend synthetic speech a natural rhythm, it does not suffice if we just find out the mean durations of pauses and lengthenings. Instead, we should rather model their durations and temporal positions in a context-sensitive way.

4. Statistical modelling of segmental durations

Because we are still seeking the most suitable statistical method to predict the durations, we carried out regression analysis on the basis of partial source material (passages from the detective story read by the actor). The input data to the statistical analysis of durations was the sequence of sounds (phonemes) and sound durations obtained by segmenting the speech wave. On the basis of a text corresponding to the speech, we formed a vector of features (with 17 features) for every sound. Those argument features were described on several hierarchical levels (phoneme, syllable, foot, word, phrase and sentence levels). We proceeded from the presumption that every sound has intrinsic duration, a vowel belongs to a concrete class of sounds (front vowels, plosive consonants, nasals etc) whose properties translate to the members of the given class; and also, that adjacent sounds impact on one another and that the duration may be influenced by both the word and the sentence structure. The output of the model or the functional feature (response) — duration — was presented as logarithmic LN (duration), because the logarithmic duration conforms more to the normal distribution. Because the argument features (explanatory variables) were numerous, an optimum selection had to be made among them, i.e. we had to locate the features most likely to affect the response.

Table 4

Summary of fit and the analysis of variance for the regression model of durations

Summary of Fit

Mean of Response	–2.7530	R-Square	0.5393
Root MSE	0.2886	Adj R-Sq	0.5368

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	26	478.5	18.403	220.87	<0.0001
Error	4906	408.8	0.0862		
C Total	4932	887.2			

The initial results of statistical modelling of multiple regression analysis revealed that the model created is statistically significant (cf. Table 4). The analysis of regression coefficients disclosed that significant features for predicting the duration of the sound were the class and length (short or long) of the current sound, the class of the next sound, the position of the sound in syllable, the position of the syllable in foot, the length of the word in feet, and the location of the word in phrase. Curiously the quantity degree of the foot, despite being the cornerstone of Estonian word prosody, was not a significant feature for prediciting the duration of a sound. Those modelling results, however, have been obtained relying on only partial data volumes. Table 5 presents the features estimated as significant by experts and the statistically significant argument features obtained by regression analysis. Acting as experts were six Estonian phoneticians. The conclusions of the experts and the results of regression analysis coincided on average to 49%.

Table 5

Expert opinions versus results of regression analysis
(ExpN — N expert, Reg — results of regression analysis,
1 — significant explanatory variable, 0 — unsignificant variable)

Explanatory variable	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6	Reg
Previous phoneme class	0	0	0	0	0	0	0
Previous phoneme length	1	1	1	1	0	0	0
Current phoneme class	1	1	1	0	1	0	1
Current phoneme length	1	1	1	1	0	0	1
Next phoneme class	1	1	0	0	0	0	1
Next phoneme length	1	0	1	1	0	0	0
Phoneme position in syllable	1	1	0	1	0	0	1
Stress of syllable	1	1	1	1	1	1	0
Type of syllable	1	0	0	1	1	1	0
Quantity degree of foot	1	1	1	1	0	1	0
Syllable position in foot	1	1	1	1	0	1	1
Length of foot in syllables	1	1	0	1	0	1	0
Foot position in word	1	0	0	1	0	1	0
Length of word in feet	1	1	0	1	0	0	1
Word position in phrase	1	1	1	1	0	0	0
Length of phrase in words	1	0	0	1	0	1	1
Length of sentence in phrases	1	0	0	0	0	0	0
Total ”correct” answers	8	11	8	7	9	7	
%	**47%**	**65%**	**47%**	**41%**	**53%**	**41%**	

Total average 49%

The analysis of prediction residuals or errors (cf. Figure 1) showed that in the distribution of errors there were three "data clusters" distanced from one another. A closer look revealed that the two right-hand clusters were constituted by pauses. The residuals may be considered, at a visual estimate, to be homoscedastic.

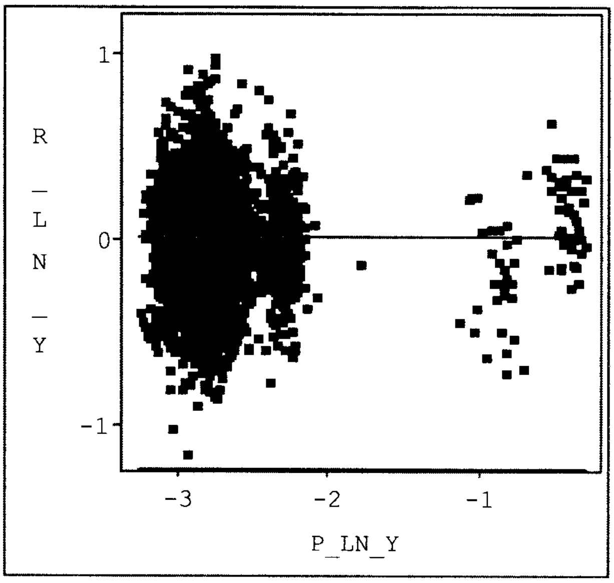

Figure 1. **Residual by predicted values of sound durations: y-axis (R_LN_Y) — prediction residuals of logarithmic sound durations, x-axis (P_LN_Y) — predicted values of logarithmic sound durations.**

5. Conclusions and future work

This paper has described the preliminary results and the first attempts to make the prosody of the output speech of a text-to-speech synthesiser of Estonian more natural. The analysis of prediction errors showed that the sounds and pauses should be handled separately at analysis. To predict the duration of sounds and pauses using statistical methods the volume of material analysed should be expanded, with various methods tested (e.g. neural networks).

REFERENCES

Campell, N. 2000, Timing in Speech. A Multilevel Process. — Prosody. Theory and Experiment, Dordrecht—Boston—London, 281—334.

Eek, A., Meister, E. 2003, Foneetilisi katseid ja arutlusi kvantiteedi alalt (I). Häälikukestusi muutvad kontekstid ja välde. — KK, 815—837.

Krull, D. 1997, Prepausal Lengthening in Estonian: Evidence from Conversational Speech. — Estonian Prosody: Papers from a Symposium. Proceedings of the International Symposium on Estonian Prosody, Tallinn, Estonia, October 29-30, 1996, Tallinn, 136—148.

Lehiste, I. 1981, Sentence and Paragraph Boundaries in Estonian. — CIFU V, Pars VI, 164—169.

Lehiste, I., Fox, R. 1993, Influence of Duration and Amplitude on the Perception of Prominence by Swedish Listeners. — Speech Communication 13, 149—154.

Mihkla, M., Meister, E., Eek A. 2000, Eesti keele tekst-kõne süntees: grafeem-foneem teisendus ja prosoodia modelleerimine. — Arvutuslingvistikalt inimesele, Tartu (Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1), 309—320.

Stout, R. 2003, Deemoni surm. CD-versioon. Loeb Andres Ots, Tallinn.

МЕЕЛИС МИХКЛА, ЮРИ КУУСИК (Таллинн)

АНАЛИЗ И МОДЕЛИРОВАНИЕ ВРЕМЕННЫХ ХАРАКТЕРИСТИК РЕЧИ ДЛЯ ЭСТОНСКОГО ТЕКСТ-РЕЧЬ-СИНТЕЗА

Текст-речь-система должна быть способной генерировать звуки и паузы продолжительностью, которая не отличалась бы значительно от подобных характеристик в обычной речи. Моделирование настоящей просодии эстонского текст-речь-синтеза базируется в основном на обобщенных результатах измерений речевых единиц изолированных слов и предложений. Поскольку выходная речь синтезатора зачастую оказывается монотонной, а слитность речи неустойчивой, широкое использование синтезатора ограничено. В данной работе сделаны первые попытки улучшить естественность выходной речи, используя для этого статистические модели слитной речи. В качестве исходного материала служили подчитка беллетристического текста профессиональным актером и чтение новостей диктором. По этим материалам исследовалась вариативность длительности пауз и предпаузное удлинение. Оказывается, если текст подчитан в нормальном темпе, то классификацию пауз вполне можно использовать в синтезе речи. Была предпринята также попытка выяснить, насколько синтаксическое членение текста связано с просодическим членением речи. С помощью регрессионного анализа выявлялись признаки, важные при прогнозировании длительности звуков, и строилась статистически оптимальная модель. Удивительно, что степень количества речевого такта, которая в эстонском языке является краеугольным камнем просодии слова, отнюдь не служит важным параметром прогнозирования длительности звуков в данном материале. Результаты моделирования сравнивались с экспертной оценкой.