MORPHOLOGICAL AND SYNTACTIC FACTORS IN PREDICTING SEGMENTAL DURATIONS FOR ESTONIAN TEXT-TO-SPEECH SYNTHESIS

Meelis Mihkla

Institute of the Estonian Language
Roosikrantsi 6, Tallinn, ESTONIA
meelis@eki.ee

ABSTRACT

Traditionally, durational models of speech units have been developed without paying much heed to morphology and part-of-speech information while predicting speech temporal structure. The aim of the present study was to find out whether the rich morphology of the Estonian language could possibly provide some additional (beside the syntactic and part-of-speech) information that could be used in predicting durations. The project is a continuation of prosody studies for Estonian text-to-speech synthesis. Sound durations in the speech of radio newsreaders were modelled by means of different statistical methods (linear regression and neural networks). Model input consisted not only of descriptors of sound context and position, but also of information on part of speech, part of sentence and morphological features. The results indicated a decrease of error in the prediction of segmental durations. Such results were in good harmony with our expectations concerning a morphologically rich language.

Keywords: morphological factors, part-of-speech, segmental durations, TTS synthesis

1. INTRODUCTION

Speech prosody being affected by very many factors and their complicated combined effects it is not easy to generate synthetic speech with a prosodically appropriate temporal structure. As a rule, morphological features are not included among the factors relevant for the temporal structure of speech (cf. [1], [5] and [6]). One reason may be that most of the studies hitherto available on text-to-speech synthesis concern languages with relatively little morphology. Finnish is different in that respect, and they, indeed, have a study on the role of morphological features on the duration of speech units [7]. As Estonian is a language with the word having a central role in grammar as well as phonetics, and with an extremely rich morphology at that, we wondered if the temporal structure of Estonian speech could possibly be affected by some morphological, lexical and maybe even syntactic features.

In some previous studies on Estonian prosody several statistical methods (linear regression, neural networks, CART) were successfully applied to predict the segmental duration of speech units [2], [3]. In the present paper the same methods were extended over certain linguistically based factors such as morphology, lexis and syntax. The linguistic knowledge used rests on available technologies prepared for the Estonian language [4], [8]. As the morphological analyser and parser already work in text-to-speech synthesis it seemed a waste not to use their information in our prosody generator. A demo version of the Estonian syntax analyser is already accessible over the Internet http://www.cs.ut.ee/~kaili/parser/demo/.

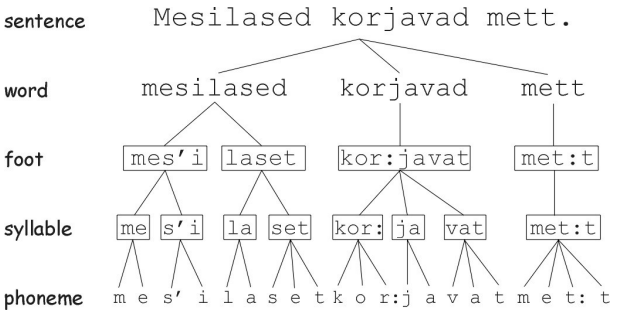
The most natural way to assess the effect of morphological, part-of-speech and syntactic factors seemed to lie through an extension of our previous methodology of statistical modelling to see how those factors may affect the functioning of durational models. The models were built on two different methods - linear regression and neural networks, the latter being a nonlinear method. The comparable effect of the factors was measured by change of output error as measured on different models.

2. INITIAL DATA AND ARGUMENT FEATURES

The initial data consisted of speech passages and news texts as pronounced by two radio newsreaders (male and female). The total amount of speech analysed included 6.3 minutes of male speech and 9.1 minutes of female speech

segmented manually into 5063 and 7010 speech sounds, respectively.

Figure 1: Hierarchical encoding of the relative position and length of a current speech unit. In this case for example the phone [l] is being estimated and its place is coded according to its position in syllable [la] of length two (phones). The syllable's position is coded in relation of foot [lased] with a length of two syllables. The foot's position is coded in relation to word [mesilased] with a length of two feet. The word is further given a code according to its place in the sentence [Mesilased korjavad mett. *'Bees gather honey.'*] with the length three words.



In durational models of Estonian sounds the argument features are represented hierarchically (Figure 1). The hierarchical levels are sentence, word, foot, syllable and phoneme. So the relative position of a speech unit, e.g. phoneme, is referred to in the hierarchical scale by describing its position in the syllable, the position of the syllable in the foot, the position of the foot in the word, and, finally, the position of the word in the sentence. In addition, as has been proved by previous studies, information on sentence and word length comes in handy.

Table 1: Input (per sound) for modelling segmental durations.

Input	*Measurement*
Left phoneme class	Nominal (9 classes)
Left phoneme length	Binary (short and long)
Current phoneme identity	Nominal (26 phonemes)
Current phoneme length	Binary
Right phoneme class	Nominal
Right phoneme length	Binary
Phoneme position in syllable	Ordinal
Syllable position in foot	Ordinal
Length of word in feet	Ordinal
Length of phrase in words	Interval
Punctuation	Binary

The phoneme segment itself is characterized by phoneme identity and phoneme length. The necessary characteristics also include the class and length of the left and right neighbours (predecessor and successor) of the current phoneme. Before supplying the morphological, part-of-speech and syntactic features the durational models were optimized, removing all but the most vital features. The aim was to reduce not only the number of features to be considered but also to avoid some possible joint effects between new and old features (see Table 1 for the features proved the most essential by our analysis).

Table 2 represents some new candidates of argument features for the input of durational models. Most of the possible features are contained in the morphological factor. The values of the morphological features and the part-of-speech information have been generated by means of the Estonian morphological analyser [8]. As can be seen, most of the morphological factors concern a certain part of speech only. Verbs, for example, are involved with the highest number of factors, whereas adverbs, adpositions and conjunctions carry a single morphological marker - invariable word. For nearly all new factors the influence is manifested on word level, except the feature "stem vs. suffix", that belongs to phoneme level. That feature was included to check whether the duration of stem sounds might differ from that of suffix sounds. The syntactic analysis of the text sentences was done manually. The response of the models consisted of logarithmed durations of sounds.

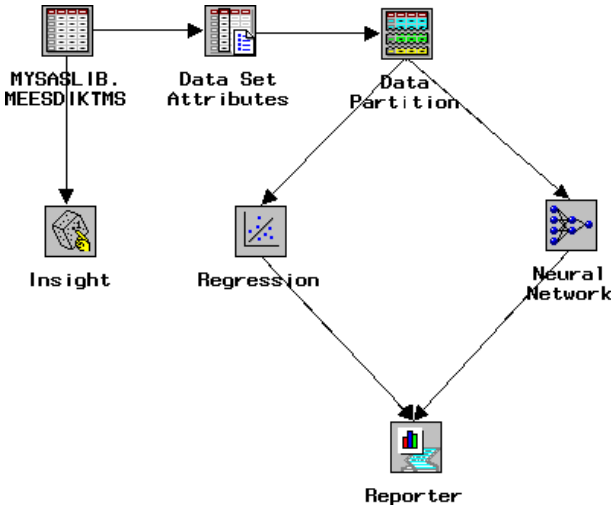
Table 2: Morphological, syntactic and lexical factors concerning the word. Number of the necessary values encoded as input to duration models.

Input	*Number of stages*
Morphological factors:	
- case	15
- number	2
- tense	2
- person	2
- voice	2
- infinit/partic	6
- stem/suffix	2
Syntactic factor:	
- part of sentence	8
Lexical factor:	
- part of speech	11

3. STATISTICAL ANALYSIS

Statistical modelling of the sounds was done by means of the SAS package 9.1. Figure 2 contains the block diagram of our data processing.

Figure 2: SAS Enterprise Miner workspace.



The methods picked were linear regression and the neural networks. The choice was purely pragmatic, enabling a comparison of the response of a linear and a non-linear model to different argument features. Linear regression used backward selection with a 0.05 significance level. The neural networks represented a multilayer perceptron with one hidden layer. For modelling purposes each newsreader's data were divided into three parts: 50% of the data were used for model training, 30% for data validation and 20% for testing.

4. RESULTS

Table 3 demonstrates a decrease of the average error in response to different information added to model input. The results indicate that, depending on the factor, part-of-speech and part-of-sentence information as well as morphological features can improve model efficiency and predictive precision by ca 0.4 – 1.3 %. Without input of morphological-syntactic information the average predictive error of segmental durations stayed within the limits of 16.5 – 18.1 %. Of morphological features separate mention has been made of stem vs. suffix. As can be seen from the table the output effect of that feature is irrelevant. Consequently, the duration of suffix sounds does not differ from that of stem sounds. The final row of Table 3 gives the total contribution of all factors to the efficiency increase of the durational model. A can be seen, the total decrease of the predictive error does not equal the sum of the effects of the individual factors, which is obviously due to co-effects occurring between some factors. The sensitivity of neural networks is raised considerably by supplying the input with the part-of-sentence (syntactic) feature. The error decreases twice as much as in the regression model. Nevertheless it is rather difficult to express the error decrease in quantitative terms. The final assessment of the effect of morphological-syntactic input is awaiting some perception tests.

Tabel 3: Results of adding morphological information, part-of-speech and part-of-sentence status to the durational models. The values represent the average decrease in error (%).

Factors	Male newsreader		Female newsreader	
	REGR	NN	REGR	NN
Part of speech	-1,02	-0,46	-1,34	-1,09
Morphology	-0,96	-0,71	-0,84	-0,86
Part of sentence	-0,37	-0,82	-0,46	-1,06
Stem vs. suffix	-0,04	-0,08	0,00	-0,03
All factors together	-1,64	-1,29	-1,61	-2,36

Table 4: The values of regression coefficients for different part-of-speech in the male and female material.

Part of speech	Male newsreader	Female newsreader
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

A far as the models analyzed are based on the speech of no more than two speakers it is certainly premature to make generalizations. Yet a visual survey of the regression coefficients suggests that the most distinct regularities concern the parameters of the part-of-speech factor. Table 4 represents the values of the regression coefficients for different parts of speech in the male and female material. Variance seems to be higher in the middle part of the table, whereas the beginning and end parts are very similar. The table reveals that the speech sounds of proper names are pronounced in average 5.22 – 6.23 ms (10 %) longer than in verbs. The average duration of sounds in newsreaders speech was 62.5 and 64.1 ms,

respectively. Nouns and adpositions were slightly prolonged. 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 numbers, however, were pronounced over 10% shorter, while pronouns and adverbs tended to be shorter by ca 5%.

5. CONCLUSION

The study was meant first and foremost as a continuation of prosody research for Estonian text-to-speech synthesis. The results revealed that addition of morphological-syntactic information to model input yields a couple of percent decrease of error in predicting segmental durations. Considering the rich morphology of the Estonian language such behaviour of the models was no surprise. As a morphological analyser is already at work in the linguistic processing of texts for Estonian speech synthesis it is obvious that some part-of-speech and morphological information will be used in duration modelling. The possible necessity of involving the rather complex syntactic analyser, however, cannot be decided without first performing some perception tests.

6. REFERENCES

[1] Campell, N. 2000. Timing in speech: a multilevel process. In M. Horne (editor), *Prosody: theory and experiment*. Dordrecht/Boston/London: Kluwer Academic Publishers, 281-334.

[2] Fishel, M., Mihkla, M. 2006. Modelling the temporal structure of newsreaders' speech on neural networks for Estonian text-to-speech synthesis. In: *Proceedings of the 11th International Conference "Speech and Computer": SPECOM2006*, St. Petersburg: Anatolya Publishers, 303 - 306.

[3] Mihkla, M., Kuusik, J. 2005. Analysis and modelling of temporal characteristics of speech for Estonian text-to-speech synthesis. *Linguistica Uralica*, XLI(2), 91 - 97.

[4] Müürisep, K. 2000. Eesti keele arvutigrammatika: süntaks. Dissertationes Mathematicae Universitatis Tartuensis 22.

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

[6] van Santen, J. 1998. Timing. In: *Multilingual text-to-speech synthesis*: The Bell Labs Approach, Sproat, R. (editor), Kluwer Academic Publishers, 115-140.

[7] Vainio, M. 2001. Artificial neural network based prosody models for Finnish text-to-speech synthesis. Helsinki: University of Helsinki.

[8] Viks, Ü. 2000. Eesti keele avatud morfoloogiamudel. -- *Arvutuslingvistikalt inimesele* (toim T. Hennoste). Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Tartu, 9-36.

7. ACKNOWLEDGEMENT

The study was financed by the national programme “Language technological support of Estonian”.