Транскрибирование, структурирование и временной анализ речевого корпуса эстонского языка при выборе единиц в системе синтеза (текст-речь)

Transcrbing, structuring and temporal analysis of fluent speech corpus for a unit selection tts system for Estonian

Meelis Mihkla (meelis@eki.ee), **Indrek Kiissel** (indrek@eki.ee),
Tõnis Nurk (tonis@eki.ee), **Liisi Piits** (liisi@eki.ee)

Institute of the Estonian Language, Tallinn, Estonia

В статье рассматриваются проблемы создания системы синтеза, основанной на выборе единиц из корпуса эстонского языка (текст-речь). Авторы предлагают правила транскрибирования и принципы фонологического структурирования, облегчающие выбор языковых единиц. Исследуется также интенсивность коллокации (сочетаемости) в зависимости от темпа речи и разрабатываются соответствующие модели длительности.

1. Background

Around the turn of the millennium (1997–2002) the first generation of Estonian corpus-based TTS synthesizer was developed (Mihkla et al. 1999). This time the basic speech units were diphones, chosen as sources of natural phone transitions. With this undertaking we also became part of the international MBROLA project. Although our first-generation synthesis was also corpus-based, its database provided no more than one diphone for each transition. It has been argued — and proved in practice — that the large number of concatenation points make the synthetic speech sound unnatural, even if the spectral discontinuities have been minimized by carefully smoothing the concatenation points, considering phonetic criteria (Donovan, Woodland 1999).

Two years ago a new project of corpus-based synthesis of Estonian was launched in the framework of the national programme "Language technological support of Estonian". The aim is to develop a high-quality second-generation speech synthesizer, based on unit selection. Our ambition is to create on the bases of moderate speech corpus (up to 60 minutes) high quality Estonian speech synthesizer for male and female voices. The new synthesizer, however, draws its acoustic material from the whole speech corpus. The idea of corpus-based or unit-selection synthesis is that the corpus is searched for maximally long phonetic strings to match the sounds to be synthesized. As compared to diphone or triphone synthesis, corpus-based speech tends to elicit considerably higher ratings of naturalness in auditory tests (Nagy et al., 2003). As the corpus in its entirety provides the acoustic basis for such synthesis, the development of an optimal corpus represents an essential task of corpus-based synthesis.

Our speech corpus (Piits et al. 2007) contains phonetically rich sentences and various phonological structures of Estonian. The corpus includes words which contain all Estonian diphones and many numerals, alongside with frequent Estonian words and expressions. Development of a unit selection based TTS system for Estonian take place in two directions: on the one hand the system is developed in the Festival environment with Cluster unit selection. On the other hand we would like to test just how far we can go by using a high-quality corpus and good algorithms of unit selection without any synthesis engine, just applying some very simple methods of signal processing. Both applications demand high quality speech corpus, thus the development of an optimal corpus represents an essential task of corpus-based synthesis. A system with a good selection module and a high-quality speech corpus may yield output speech of extremely high quality, even if the signal processing module is rather simple (Bozkurt et al., 2002).

In order to convert an Estonian written text into synthesized speech we have to solve the following

tasks: to convert an orthographic text into a phonetic-phonological one and to compile rules or models for the control of segment durations and F0 contours. The most difficult problem is: how to find the third quantity degree Q3 and palatalization automatically in the written text. We also investigate how collocation strength might correlate with speech rate and what role it might play in duration models. In order to optimize the unit selection algorithm and to guarantee the necessary quality of the synthetic speech the whole speech database as well as the utterance to be synthesized is represented as a phonological tree.

2. Transcription rules for fluent speech corpus

When transforming an orthographic Estonian text into a pronounced text one should take into account that not every phonological opposition is spelt out in Estonian. This applies, for example, palatalized vs. unpalatalized consonants and, in a general case, to the 2nd and 3rd quantity degrees. In Estonian long stressed syllables can be pronounced with (third quantity degree: Q3) or without (second quantity degree: Q2) prosodical quantity. In principle, both palatalization and word quantity are marked in the lexicon, after morphological analysis of the sentence. Still, there are cases where the phonetic quantity and palatalization defined on the basis of the lexicon need additional adaptation to meet the rules of fluent speech.

In Estonian, which is considered a stress-timing language, the foot (1–3 syllables) carries such prosodic phenomena as stress and quantity. Two- or three-syllable words may be classified into first-, second- and third-quantity words, whereas the mono-syllabics have, theoretically, all been marked out for the third quantity. Actually fluent speech is characterized by a considerable number of monosyllabic words (pronouns, conjunctions, adverbs) capable of either adhering to other words in longer feet or occurring in an unstressed position as clitics in the speech flow (Hint 1998:145–146; Lehiste 1997:11). Therefore the more frequent pronouns and conjunctions are not subjected to the general rules of transcription, but treated as special cases, e.g the conjunctions *et* 'so as to' [*ett*, not *et:t*], *kui* 'if; than' [*kui*, not *kui:*]; the pronoun *neid* 'he/she Part. Pl.' [*neit*, not *nei:t*]. At the same time, the third quantity is retained in content words of a similar structure, e.g. *vett* 'water Part. Sg.' [*vet:t*].

The Estonian language has four palatalized speech sounds: the palatal lateral approximant [l'], the palatal nasal [n'], the alveolar ejective fricative [s'], and the voiceless palatal plosive [t']. Our orthographic spelling makes no difference between the palatalized and unpalatalized sounds, even though there are several cases where palatalization does have a distinctive function as, for example, in *palk* 'salary' [*palk:k*] and *palk* 'log' [*pal'k:k*]. In addition to the lexicon-based palatalization, palatalization is a feature added automatically to such *l, n, s, t* that, preceding *i* or *j,* immediatedly follow either the vowel of a main stress syllable or a palatalised consonant, e.g. *hiilima* 'to sneak' [*hii:lima -> hii:l'ima*], *kasti* 'box Gen. Sg.' [*kas'ti -> kas't'i*], but *kunsti* 'art Gen. Sg.' [*kun'sti*].

Foreign letters are transcribed according to the Estonian tradition: the fricatives *z* and *ž* (voiced retroflex fricative), for example, sound voiced in many languages, but not in Estonian, where their pronunciation does not differ from that of *s* and *š* (voiceless retroflex fricative).

The long *üü* (close front round vowel) is diphthongized if followed by a vowel or *j,* like, e.g., in *müüjad* 'salesperson Nom. Pl.' [*myi:jat*], *hüüe* 'shout' [*hyie*]. If a vowel follows a long *ii* or an *i*-final diphthong, it is pronounced as if preceded by the glide *j*, e.g. *saiu* 'white bread Part. Pl.' [*sai:ju*]. A similar rule works for a long *uu* and an *u*-final diphthong, where the *u* is pronounced as if followed by the voiced labiovelar approximant *w*, e.g. *suue* '(river) mouth' [*suuwe*].

If three stops happen to meet on a word boundary, the single plosive phonemes are usually reduced to such an extent that, as a rule, a dissimilatory loss can be diagnosed, e.g. the sequence *kõlblik kokk* 'a fit cook' is transcribed as *kõl:plik kok:k*, not as *kõl:plikk kok:k*. Dissimilatory loss may also be found in vowels, e.g. the phrase *ma ei tea* 'I don't know' is pronounced as [*maitea*].

The letters *k, p, t, t', f* and *š* occurring between voiced sounds on syllable boundary or at the end of a word are transcribed doubly, whereas in the immediate neighbourhood of a voiceless sound as well as at the beginning of a word they remain single, e.g. *kokpit* 'cockpit' [*kokpitt*], *Aafrika* 'Africa' [*aaffrikka*], *part* 'duck' [*part:t*], *kašelott* 'cachalot' [*kaššelot':t'*].

3. Structuring the speech corpus

In order to optimize the unit selection algorithm and to guarantee the necessary quality of the synthetic speech the whole speech database as well as the utterance to be synthesized is represented as a phonological tree. Figure 1 represents a fragment of the database tree. Our phonological tree has the following levels: phoneme, syllable structure, syllable, foot, word, phrase, and sentence (cf. Breen et al. 1998, Taylor et al. 1999). Search for appropriate speech units involves all those levels, beginning from the higher ones, e. g. preferring longer units.

The free word order and close intertwinement of the Estonian syntactic phrases frustrates, as a rule, the attempts to build a binary tree between the word and sentence levels. For the present project a phrase is defined as a clause that is separated by an intrasentence punctuation mark or conjunction and that includes a predicate.

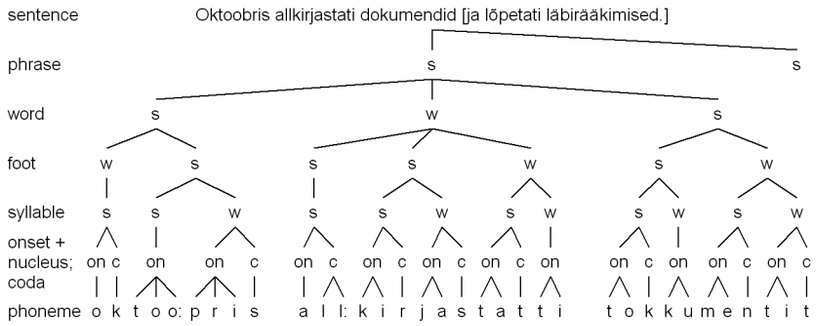

Figure 1. Fragment of the phonological tree

Black and Taylor's phonological structure has been criticized for fixed representation of word boundaries (Möbius 2000). True, there are some Estonian (compound) words (e.g. *raudtee* 'railway', *õunapuu* 'apple-tree') with co-articulation occurring on word boundary, yet the basic unit in Estonian speech is the word and thus an emphasis on the marking of word boundaries presents no problem for speech synthesis. But there is an extremely rich morphology to consider. Thus we first attempted a joint use of the phonological tree and a morphological one consisting of word forms divided into stem and grammatical morphemes. Unfortunately the representation of phonological and morphological information in one and the same tree turned out too complicated a task, because morpheme and syllable boundaries do not coincide. So we had to keep to the phonological tree after all. Still, our consideration for morphological information had been not quite in vain, providing great help in finding phonologically suitable units.

In some ways the branching of the phonological tree on sub-syllable levels is also morphologically bound. Syllable division starts from separating the coda. This enables the use of the onset and nucleus in word form synthesis even in case the coda is wrong. Thus, after separating the coda we can use the word form *ka-la-l* 'fish Adess. Sg.' to form, e.g. *ka-la-s* 'fish Iness. Sg.' *ka-la-st* 'fish Elat. Sg.' or *ka-la-lt* 'fish Ablat. Sg.'

4. Effect of collocational strength on speech rate

While recording an Estonian corpus for corpus-based synthesis some fluctuations of the speech rate were observed, even though the text was read out by a professional radio announcer. The slowings down could be due to difficult clusters (the corpus was required to contain all diphones possible in Estonian, however rare (see Piits et al. 2007), which could, in turn, occur in rare words. A quickening rate, however, could have to do with frequent words as well as collocational phrases. It has been argued before that the high frequency of a word and the predictability of its context may have a reducing effect on the pronunciation of the word (Pluymaekers et al. 2005, Bell et al. 2003). In some cases the effects of word frequency and contextual predictability on word duration have been studied in combination (M. L. Gregory et al.1999).

In Estonian, the word has a very important role both in grammar and phonetics, while the morphology is extremely rich. The aim of the present study is to find out if, apart from word frequency, Estonian word length could in any way depend on the collocational strength between the words.

We set up the following hypothesis: collocational strength between words in Estonian has an effect on word duration, i.e. words occurring side by side more often tend to be pronounced more rapidly. Our scrutiny is focused on the verb *olema* 'be' as the most frequent word in Estonian.

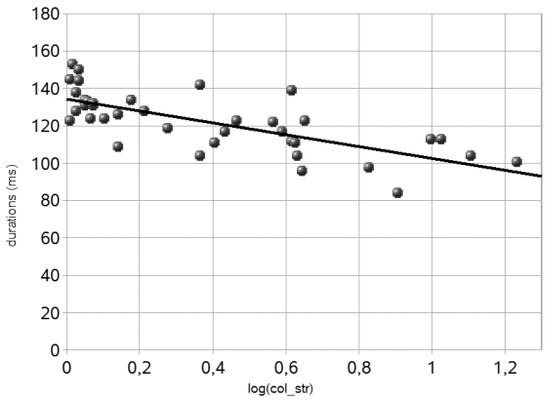

Figure 2. Relation between duration of stem sounds and collocational strength.

We investigated how collocation strength might correlate with the durations of *olema* forms and what role it might play in predictive models. Figure 2 shows the relation between duration of stem sounds of verb *olema* 'be' and collocational strength (in logarithmic scale). The hypothesis was tested by means of different statistical methods (linear regression, CART trees), enabling to disclose small, hidden, but possibly significant effects between input and output (Sagisaka 2003).

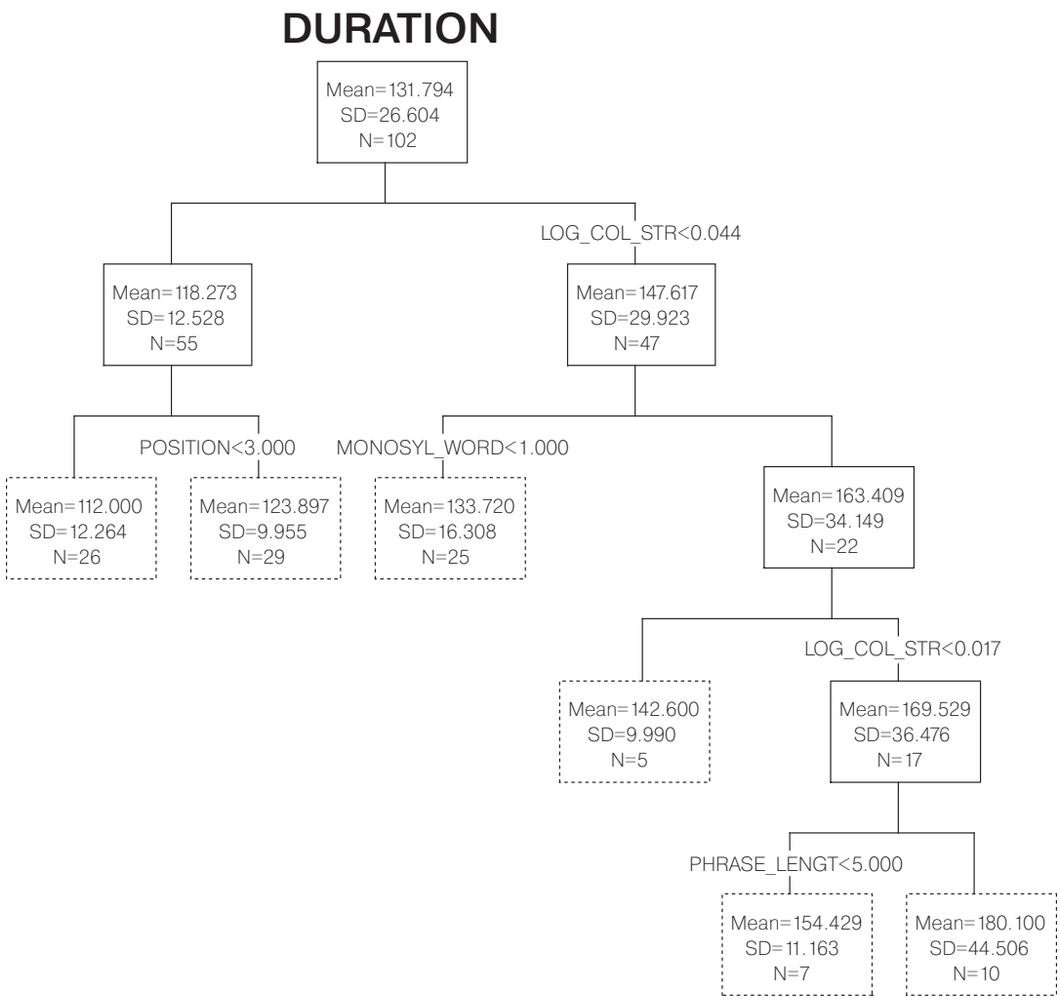

Figure 3. CART model of durations of the verb ***olema*** 'be'.

A simple durational model was compiled to predict the duration of the verb *olema* 'be' from collocation strength, length of phrase, position of the verb in the phrase, and a binary characteristic indicating whether a concrete verb form had just one syllable or more (Fig. 3). According to the resulting models there were two features — the binary one and collocational strength — that were significant in all models. Consequently, in the material studied collocational strength does have an effect on the durations of the verb *olema* 'be'.

5. Conclusion

The aim of the speech corpus described was to develop an acoustic basis for a relatively naturally sounding synthetic speech. To reduce the number of concatenation points in the synthetic utterance it was necessary to create a speech corpus enabling searching for units larger than diphones. The phonological structure describes different levels from where the units were found. The rules for transforming an orthographic Estonian text into a pronounced text should certainly provide for the perception of phonologically essential distinctions. Yet apart from that the transcription rules should not overlook some additional features, which, without being distinctive, still play an important role in making fluent speech sound natural. The study also demonstrated that the strength of collocation between words shortens the duration of the Estonian verb *olema* 'be' and that contextual predictability is a significant feature to be considered in developing models of word duration. Whether this indicates a stable relation between input and output or an occasional hidden one is a question pending further research involving measurement of collocation strength and durations of other words on more copious speech material.

Acknowledgements

This work is supported by National Programme for Estonian Language Technology, grant ETF7998 and project SF0050023s09.

References

1. *Bell A., Jurafsky D., Fosler-Lussier E., Girand C., Gregory M. and Gildea D.* 2003. Effects of disfluencies, predictability and utterance position on word form variation in English conversation. // Journal of the Acoustical Society of America, 113(2), pp. 1001–1024.
2. *Bozkurt B., Dutoit, T., Prudon, R. C., D'Alessandro, C., Pagel, V.* 2004. Reducing discontinuities at synthesis time for corpus-based speech synthesis. // In: Narayanan, S.; Alwan A. (eds). Text To Speech Synthesis: New Paradigms and Advances, pp. 1–17.
3. *Breen, A. P., Jackson, P.* 1998. Non-Uniform unit selection and the similarity metric within BT's Laureate TTS system. // In: Proc. Third ESCA Workshop on Speech Synthesis, pp. 373–376.
4. *Donovan, R. E., Woodland P. C.* 1999. A hidden Markov-model-based trainable speech synthesizer. // Computer Speech and Language 13, pp. 223–241.
5. *Gregory M. L., Raymond W. D., Bell A., Fosler-Lussier E. and Jurafsky D.* 1999. The effects of collocational strength and contextual predictability in lexical production. // CLS-99, pp. 151–166. Chicago: University of Chicago.
6. *Hint M.* 1998. Häälikutest sõnadeni. // Eesti Keele Sihtasutus, Tallinn, pp. 145–146.
7. *Lehiste, I.* 1997. Search in phonetic correlates in Estonian prosody. // In: Lehiste, I., Ross, J. (eds.) Estonian Prosody: Papers from Symposium, Institute of the Estonian Language, Tallinn, pp. 11–35.
8. *Mihkla, M., Eek, A., Meister, E.* 1999. Diphone synthesis of Estonian. // In: Dialogue'99 : Computational Linguistics and its Applications: International Workshop: Proceedings. Vol. 2. Applications. (Toim.) Narin'yani,A.S.. Tarusa: 1999, pp. 351–353.
9. *Möbius, B.* 2000. Corpus-based speech synthesis: methods and challenges. // Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart) 6(4), pp. 87–116.
10. *Nagy, A., Pesti, P., Németh, G., Bőhm, T.* 2005. Design Issues of a Corpus-Based Speech Synthesizer. // Hungarian Journal on Communications 6, pp. 18–24.
11. *Piits L., Mihkla M., Nurk T. and Kiissel I.* 2007. Designing a speech corpus for Estonian unit selection synthesis. // Nodalida 2007 Proceedings: The 16th Nordic Conference of Computational Linguistics, pp. 367–371.
12. *Pluymaekers M., Ernestus M. and Baayen H. R.* 2005. Articulatory planning is continuous and sensitive to informational redundancy. // Phonetica, 62, pp. 146–159.
13. *Sagisaka Y.* 2003. Modelling and perception of temporal characteristics in speech. // Proceedings of 15th International Congress of Phonetic Sciences, Barcelona, pp. 1–6.
14. *Taylor, P., Black, A. W.* 1999. Speech synthesis by phonological structure matching. // Proc. Eurospeech'99 Budapest, pp. 623–626.