Designing a Speech Corpus for Estonian Unit Selection Synthesis

Liisi Piits
Institute of the Estonian Language
Roosikrantsi 6, Tallinn 10119, Estonia
liisi@eki.ee

Meelis Mihkla
Institute of the Estonian Language
Roosikrantsi 6, Tallinn 10119, Estonia
meelis@eki.ee

Tõnis Nurk
Institute of the Estonian Language
Roosikrantsi 6, Tallinn 10119, Estonia
tonis@eki.ee

Indrek Kiissel
Institute of the Estonian Language
Roosikrantsi 6, Tallinn 10119, Estonia
indrek@eki.ee

Abstract

The article reports the development of a speech corpus for Estonian text-to-speech synthesis based on unit selection. Introduced are the principles of the corpus as well as the procedure of its creation, from text compilation to corpus analysis and text recording. Also described are the choices made in the process of producing a text of 400 sentences, the relevant lexical and morphological preferences, and the way to the most natural sentence context for the words used.

1 Introduction

Text-to-speech synthesis means that synthetic speech is automatically generated from a written text. The understandability and naturalness of output speech depends on linguistic preprocessing of the input text, the prosody generator, signal processing and the quality of the speech database used. 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 and Woodland 1999). 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. 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).

Considering an optimal database for Estonian text-to-speech synthesis it should obviously contain phonetically rich sentences and different phonological structures of Estonian. The corpus words should also include all Estonian diphones. The first database for Estonian speech synthesis consisted of ca 1700 diphones (Mihkla et al., 1998). The experience accumulated during the creation of that database came in handy while developing our corpus. Aim was a speech corpus that would not be too big (up to 60 minutes), yet representative enough from phonetic and phonological aspects, containing many numbers and years, alongside with frequent Estonian words and expressions. Even though a necessity for repeat recordings to complement the corpus cannot be ruled out entirely, material should serve for synthesis of an arbitrary text as well as for limited domain applications.

2 Text corpus development

The first decision to be made concerned the size of the corpus. This meant a compromise between the minimum and maximum sizes. A maximum size would mean a greater probability of the corpus containing the biggest possible units to match the text to be synthesized- from sound strings to words or even phrases. Unfortunately, big databases have

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 367–371

been found complicated to maintain and even more complicated to annotate (Breen and Jackson, 1998). Moreover, segmentation and tagging of corpus units is a cumbersome and time-consuming process - it has been found that a one-minute corpus takes 1000 minutes to mark up (Mihkla et al., 1998). This is why we decided to make the corpus as small as possible, yet containing as much relevant material as possible.

2.1 Stage 1: diphones

It was decided that the smallest searchable unit of the corpus would be a diphone. Therefore it was important to ensure that the corpus contains all diphones possible in Estonian. We already had a word list, compiled for an earlier synthesizer based on diphone selection, featuring all diphones occurring in Estonian (Mihkla et al., 1998). Most of those words were taken as the basis for the new corpus. However, as the words had not been included in the list in their natural sentence context, our first task was to provide a sentence context for them.

So, Stage 1 of the corpus development started with combining the list words to make meaningful sentences. During that process one had to keep a watchful eye on the pronounceability of words and sentences, considering sentence length as well as word structure. Both too long and too short sentences were to be avoided. For English sentences it has been argued that too short sentences (less than 5 words) have a deviant prosody, while too long ones (more than 15 words) tend to elicit more mistakes when read out (Kominek and Black, 2003). As Estonian is a more synthetic language than English we did not stick to the five-word limit, ending up at seven words in an average sentence.

Among the sound combinations and syllable types of a natural language there are some that are easy to pronounce and some others that are not. The latter, being more demanding on speech organs, are used less frequently. Such sound sequences and syllable types are called marked ones (Hint, 1998). As the word list was meant to include all diphones possible in Estonian it contained not only frequent words but also the rare words with marked structure. There were even some diphones not allowed by Estonian phonotactics, but they had to be included because of their occurrence in foreign words. In addition, the list contained some nonsense words with sound combinations theoretically allowed by the rules of Estonian word structure, yet not realized. For example, the diphone *üf* can be found in the 2nd quantity degree, as in the loanword *küfoos* 'kyphosis', but for the 3rd-quantity diphone *üf*: a nonsense word **süf:fi* had to be made up, as the theoretically possible diphone cannot be derived from its 2nd-quantity equivalent. The number of such nonsense strings included in the corpus was 18.

In sentences we generally tried to disperse the words with a marked structure among the unmarked ones. Most of the nonsense strings, however, were given a concentrated presentation in special sentences: e.g. **Puls:s *kõõ:l'is seda *võõ:ba ehk* mõõ:du.*

At the end of Stage 1 the corpus contained 178 sentences, with 1244 words all told.

2.2 Stage 2: words and phrases

While diphone is a minimal unit of the corpus, a word or even a phrase is seen as a unit of maximal length. The aim of stage 2 was to supply sentences containing the most frequent Estonian words and phrases. As the synthesizer is meant for texts without domain limitations the corpus vocabulary was to cover a wide selection of spheres. The words were selected from *Frequency Dictionary of Standard Estonian* (Kaalep and Muischnek, 2002), which is based on texts from media and fiction.

The aim was to make an addition of 1000 most frequent words. Frequency measurement is complicated due to Estonian morphology. The numerous cases of stem alternation and agglutination are the reason why a word may have many forms. A noun, for example, may yield 28 word forms, each of a different grammatical meaning. As generally most of the forms are made up of a word stem and various grammatical morphemes it was found sufficient to include just the stem of a high-frequency word, which could take certain grammatical morphemes also present in the corpus. The word forms contained in the corpus were to cover all paradigms of declinable as well as conjugable words. Also, the formative variants containing different allomorphs were to be represented.

Besides agglutinative formation there was inflection to be reckoned with. In Estonian, meaning can also be conveyed by stem alternation. Gradational words have at least two stem variants, both taking different grammatical markers and endings. Therefore, different stem variants also needed to be

included, if at all possible, e.g. ***haka****-ta* 'to begin' and ***hakka****-b* *'*begins', ***mees*** 'man Nom. Sg.' and ***mehe*** 'man Gen. Sg.'**,** ***krooni*** 'crown Gen. Sg.' and ***kroo:ni*** 'crown. Part. Sg'.

Besides grammatical markers and endings the corpus was provided with words containing the most productive derivational suffixes like in: *moodusta****mine*** *'*formation', *must****lanna*** 'gypsy woman', *võist****kond*** 'team', *raha****ndus*** 'finance' etc.

The nominative and genitive forms of cardinal and ordinal numerals were also included, and the most frequent place names - not only Estonian ones, but also some foreign toponyms salient in the Estonian cultural context, such as *Soome* *'*Finland', *Rootsi* 'Sweden', *Venemaa* 'Russia', *Läti* *'*Latvia', *Ameerika* 'America', *Saksamaa 'Germany'* etc.

While constructing sentences we always aimed at finding the most natural context for the words to be included. To find the normal sentence context of the words we used a corpus portal developed at the University of Leipzig[1] to compute the most frequent left and right collocations from a large database. This provided the background for sentence compilation.

The final corpus consisted of 400 sentences with 2811 words.

3 Corpus analysis

In parallel with corpus compilation a constant process of analysis was going on to diagnose its possible weak points. First we had to find out if all Estonian diphones, existing as well as theoretical, are present in the corpus and with what frequency.

		m	n	n’	o	p	r	s	
#		141	79		105	205	71	171	
a		3	55	13	9	86	52	140	
e		35	52	14	4	15	48	19	
f			1		11		3	2	
h		2	3		12		1		
i	...	34	80	15	4	18	17	174	...
j					10				
k		4	3		85	2	20	71	
l		14	8		19	6	1	4	
l’		2			lo	2		2	
m		31	8		11	10	1	6	
					...				

Table 1. Frequency of diphone occurrence at an intermediate stage of corpus compilation

The gray cells stand for diphones missing from the corpus, however possible theoretically. As Estonian is fond of active compounding we had also to consider such sound combinations as might emerge at compound boundary. So some additional words (mostly compounds) were found to fill in the gray cells. The crossed cells stand for diphones theoretically impossible in Estonian, or at least impossible to find by the means available.

As our corpus was meant to be as rich as possible phonetically and phonologically we had to include many sound combinations that were less frequent, yet vital for TTS. As can be seen in Figure 1 the synthesis corpus has considerably more of rare phonemes than the mixed corpus of Estonian[2], although in general there is no significant difference between the phoneme frequencies across the two corpora.

4 Recording

The main criterion in voice donor selection was their ability to read out the whole text at relatively constant prosodic parameters. As a result, a professional radio announcer (female) was chosen. The recording (sampling frequency: 44.1 KHz, resolution: 16 bits) was made at a studio of the Estonian Radio. The recording session lasted about an hour, yielding 51 minutes of recorded speech. The text was read out relatively monotonously, as the pitch amplitude was to be kept relatively low. The reason is that although the pitch of the synthetic signal is later subjected to modification by some signal processing methods, a large-scale interference is bound to have an undesirable effect on the quality of synthesis.

The recording pursued canonical Estonian pronunciation. It was based on the pronunciation received by *Estonian orthological dictionary* (ÕS 2006) as the would-be source of diacritics to aid text synthesis. Problems ensued from the word-final *s, n, t,* and *l*, as the orthological dictionary requires their palatalization in some positions (*roos’* 'rose', *geen’* 'gene'), although their palatalization has ceased to be consistent in modern Estonian. Even though no studies have been conducted to prove it, it seems that there is a tendency to use the non-palatalized variants in those positions.

[1] http://corpora.informatik.uni_leipzig-de/

[2] http://www.cl.ut.ee/korpused/segakorpus

Therefore those cases were recorded relying on the speaker's pronunciation.

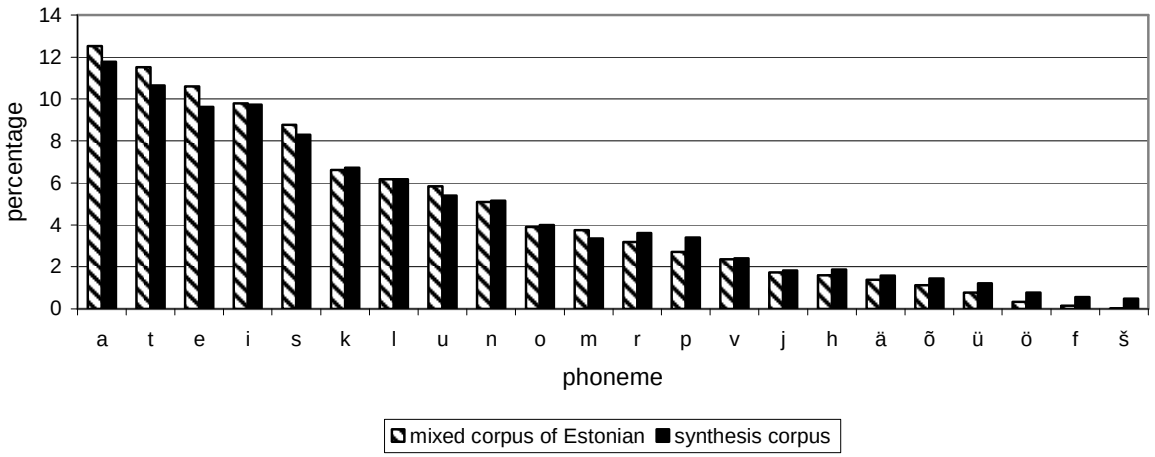

Figure 1. Frequency of phoneme occurrence in a mixed corpus of Estonian vs. the synthesis corpus.

Problems were also caused by some occasional fluctuations of the speech rate. Some of the corpus sentences included rare diphones, which may occur only in words extremely rare in Estonian, or in artificial compounds. This caused the speech rate to drop as compared to the sentences with frequent words in their normal contexts of occurrence. Whether and to what extent such fluctuations in speech rate may affect the quality of the synthesis will be revealed in the practical use of the synthesizer, which is also the proof of a necessity for additions to the corpus and for repeat recordings.

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 article provides specifics on the material included in the 400-sentence corpus. Corpus development being the first step towards natural-like synthetic speech, we are now busy tagging and segmenting the speech material and laying a phonological structure on the speech corpus.

Acknowledgement

The support from the program Language Technology Support of the Estonian has made the present work possible.

References

Baris Bozkurt, Thierry Dutoit, Romain C. Prudon, Christophe D'Alessandro and Vincent Pagel. 2004. Reducing discontinuities at synthesis time for corpus-based speech synthesis, in *Text To Speech Synthesis: New Paradigms and Advances,* (S. Narayanan, A. Alwan, ed.), Prentice Hall PTR.

Andy P. Breen and Peter Jackson. 1998. Non-Uniform Unit Selection and the Similarity Metric within BT's Laureate TTS System. *Proceedings of the Third ESCA Workshop on Speech Synthesis.*

Robert E. Donovan and Phil C. Woodland. 1999. A hidden Markov-model-based trainable speech synthesizer. *Computer Speech and language,* 13:223-241.

Mati Hint. 1998. *Häälikutest sõnadeni. Eesti keele häälikusüsteem üldkeeleteaduslikul taustal.* Tallinn.

Heiki-Jaan Kaalep and Kadri Muischnek. 2002. *Eesti kirjakeele sagedussõnastik,* Tartu.

John Kominek and Alan W. Black. 2003. *CMU ARCTIC databases for speech synthesis.* Carnegie Mellon University.

Meelis Mihkla, Arvo Eek and Einar Meister. 1998. Creation of the Estonian Diphone Database for Text-

to-Speech Synthesis. *Linguistica Uralica*, 34(3):334-340.

András Nagy, Péter Pesti, Géza Németh and Tamás Bőhm. 2005. Design Issues of a Corpus-Based Speech Synthesizer, *Hungarian Journal on Communications*, 6:18-24.

ÕS 2006 – *Eesti õigekeelsussõnaraamat 2006*. Eesti Keele Sihtasutus, Tallinn.