СИНТЕЗ РЕЧИ ЭСТОНСКОГО ЯЗЫКА: ПРИМЕНЕНИЕ И ВЫЗОВЫ

В XXI веке развитие эстонского синтезатора речи осуществлялось на основе наиболее распространенных методов и находящихся в свободном доступе программных ресурсов (MBROLA, Festival, eSpeak, HTS). В области применения синтезатора речи особое внимание уделялось потребностям людей, страдающим нарушениями зрения (система зачитывания электронных текстов, озвучивание субтитров, генерирование аудиокниг). Проблема достижения естественности звучания и его экспрессивности является важнейшим вызовом для эстоноязычного синтезатора речи. В работе рассматриваются вопросы моделирования речевой просодии посредством статистических методов и взаимосвязи просодии с другими языковыми уровнями и внеязыковыми факторами. Результаты анализа оказывающих эмоциональное воздействие акустических средств (пауз, темпа речи, формантов, интенсивности и основного тона) создают предпосылки для моделирования эмоций на уровне синтезатора речи. Рассматриваются также интерфейсы синтезатора речи, позволяющие управлять процессом речевого синтеза, отслеживать процесс видоизменения текста-речи, принимать во внимание структуру передаваемого текста и изменять параметры синтезируемых звуков (громкость звучания, темп речи, высота звучания голоса) на уровне различных голосовых решений.

ESTONIAN SPEECH SYNTHESIS: APPLICATIONS AND CHALLENGES

Mihkla M. (meelis@eki.ee),
Hein I. (kiisu@eki.ee),
Kalvik M. (mariliis@eki.ee),
Kiissel I. (indrek@eki.ee),
Tamuri K. (kairi.tamuri@eki.ee)
Institute of the Estonian Language, Tallinn, Estonia

Sirts R. (risto.sirts@err.ee)
Estonian Public Broadcasting, Tallinn, Estonia

In the 21st century Estonian speech synthesis has been developed using the more widespread methods and freeware development systems (MBROLA, Festival, eSpeak, HTS). The applications have hitherto been developed mainly in view of the needs of the visually impaired (audio system for reading electronic texts, voicing of subtitles, creation of audiobooks). The major challenges currently facing the Estonian specialists are naturalness of the output speech and expressive speech synthesis. The article is concerned with the issues of statistical modelling of the prosody of synthesized speech and the relations of prosody with other language levels as well as with extralinguistic features. Analysis of the emotion-bound acoustic parameters (pauses, speech rate, formants, intensity and pitch) enable one to model emotions for speech synthesis. In addition, speech synthesis interfaces are discussed. By means of such interfaces users could control the process of speech synthesis, monitor text-to speech transformation, follow text structure and vary the parameters (voice loudness, speech rate, voice pitch) of the synthetic voice in various voice applications.

Key words: speech synthesis, voicing of subtitles, emotional speech, Estonian

1. Introduction

The history of speech synthesis knows various TTS (text-to-speech) applications from reading systems for the blind to modern interactive dialogue systems. Application of the first synthesizers was mainly limited by the machine-like sound and poor intelligibility of the output speech. Moreover, the past half century of history has witnessed several misinterpretations of the essence of speech synthesis and TTS applications. Thinking back to the 1980s with the conferences on automatic recognition of auditory images (автоматическое распознавание слуховых образов — АРСО), it was often complained that the advent of speech synthesizers failed to meet the enthusiastic expectations of the scientific and technological revolution; for many people it rather looked like a fiasco. This was not mainly because the first synthesizers sounded like robots and were sometimes hard to understand, but most of the disappointment was due to the fact that the synthesizer could not understand human speech. Somehow the sci-fi novels and films had inculcated the idea that a speaking machine should also recognize and understand speech. Doubts were even expressed whether speech synthesis was necessary at all. Today, however, TTS output has become so good that sci-fi film producers have to make a special effort to supply synthetic speech with some additional machine-like characteristics in order to prevent the audience from thinking the robots can understand speech (Taylor 2009). Despite such misinterpretations of the abilities of a speech synthesizer and undue attribution of features, modern TTS systems have quite a lot of applications as well as challenges to face, while speech synthesis contributes, first and foremost, to technologies of speech recognition and understanding (artificial intellect); however, people often find it hard to think of the two speech technological components as separate.

In Estonia such misinterpretation of speech technological abilities and overestimation of speech synthesis has not been a problem. If a language has just a million speakers, its speech technological applications are not too profitable and the problem often is how to find enough users for the applications. The Estonian experience of speech synthesis has a history of over forty years now. Already the advent of the first generation systems was accompanied by active efforts to create user applications, in particular for the visually impaired (Remmel, Tago 1984). About 15 years ago, after a short break, research and development into Estonian speech synthesis was renewed and has been advancing briskly ever since, using various methods. Speech synthesis being one of the vital language technological tools providing for the survival and versatile development of any national language in the 21st century information society, R&D into speech synthesis has enjoyed considerable state support in the recent years.

The article will survey the available Estonian speech synthesizers and their applications (e. g. an audio system for reading electronic texts, the voicing of subtitles and creation of audiobooks). The major challenges currently facing the Estonian specialists are naturalness of the output speech and expressive speech synthesis. The article is concerned with the issues of statistical modelling the prosody of synthesized speech and the relations of prosody with other language levels as well as with extralinguistic features. On the basis of the Estonian Emotional Speech Corpus the acoustic characteristics (pauses, speech rate, formants, intensity, pitch) of some basic emotions (anger, joy, sadness) and of neutral speech are being analysed. The aim is to ascertain the emotional and emotion-specific acoustic parameters enabling one to model emotions for synthetic speech. Other notable applications include smart interfaces (Speech Application Programming Interface), enabling one to control the process of speech synthesis, to monitor text-to speech transformation, to follow text structure and vary the parameters (voice loudness, speech rate, voice pitch) of the synthetic voice in various voice applications.

2. Modules of Estonian speech synthesis

The first Estonian speech synthesizers were parametric, mainly using formant-based synthesis. Over the past decade the research has concentrated on compilation-based synthesis using donor voices, without, however, completely neglecting the formants. As the local specialists are very few indeed most of our products use ready-made development systems, while he local efforts are mainly focused on language-specific issues.

2.1. Diphone speech synthesizers

Around the turn of the millennium an Estonian concatenative text-to-speech synthesizer was produced (Mihkla, Eek, Meister 1999). The speech units used were diphones drawn from a 1900-strong database of sound-to-sound transitions. Figure 1 represents a block diagram of an Estonian diphone synthesizer, explaining what building blocks or modules are necessary for producing synthetic speech from any text written in Estonian. In principle, the same scheme underlies the TTS conversion with other methods of synthesis as well. Most of the procedures, rules, formalisms, lexicons and databases are language-specific modules, except, at least in a general case, the (double-framed) block of signal processing. The diphone synthesis was performed using the MBROLA engine. Many people with a visual impairment still use the Estonian diphone-based synthesizer. The main shortcomings of MBROLA synthesizers are due to neglect of further development of the project, which, considering the recent developments in hardware is causing compatibility problems (e. g. the ghost echo in the case of multi-kernel processors).

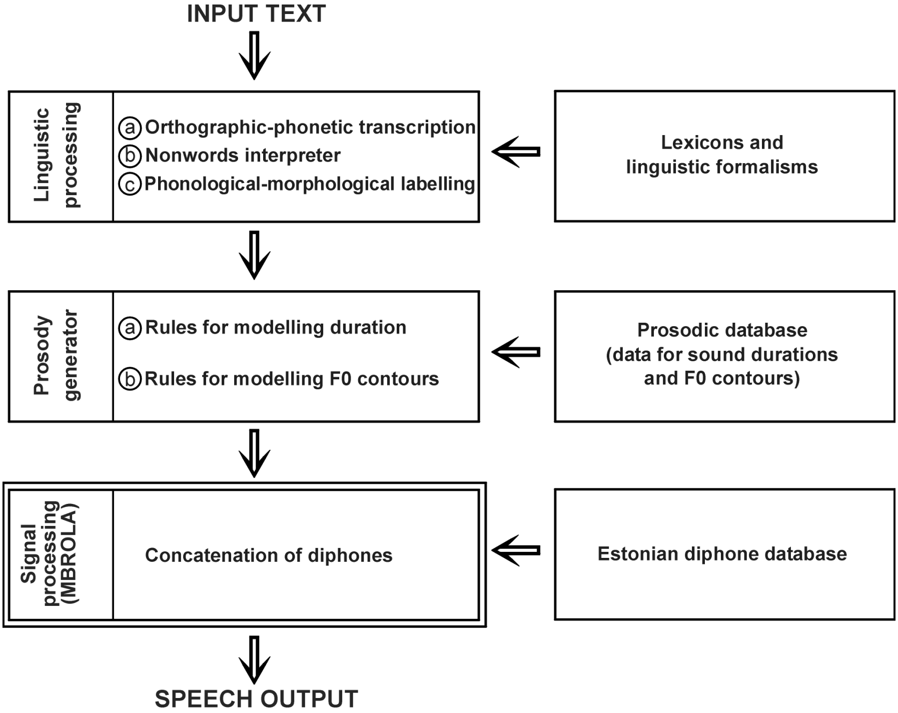
INPUT TEXT
Linguistic processing
ⓐ Orthographic-phonetic transcription
ⓑ Nonwords interpreter
ⓒ Phonological-morphological labelling
Lexicons and linguistic formalisms
Prosody generator
ⓐ Rules for modelling duration
ⓑ Rules for modelling F0 contours
Prosodic database (data for sound durations and F0 contours)
Signal processing (MBROLA)
Concatenation of diphones
Estonian diphone database
SPEECH OUTPUT

Figure 1. Block diagram of an Estonian diphone synthesizer

2.2. Synthetic voices based on unit selection

Whereas the diphone database used in the first generation compilation-based synthesis contained precisely one diphone for each sound-to-sound transition, speech synthesis based on unit selection draws from acoustic material of the whole speech corpus. The creation of a synthesizer using corpus-based unit selection is a process consisting of the following six stages:

- creation of a text corpus;
- reader selection and the storing of speech material;
- tagging of the corpus material and its segmentation into speech units;
- establishing the sound system and choice of parameters;
- morphological text analysis;
- selection of speech units for synthesis.

The text corpus needs to be sufficiently representative both from the phonetic and phonological aspects, containing all possible diphones, a great deal of numbers and years as well as frequent Estonian words and fixed combinations (Piits et al. 2007). Our minimal text corpus contains 400 sentences while the bigger ones have up to 4500. In the beginning we only used readers with trained voices (radio announcers) expected to be able to present the whole text with relatively stable prosodic parameters. However, as some of the announcers tended to display excessive expressiveness we also recorded the reading of ordinary people and in the end the synthesis based on the material of their untrained voices turned out quite satisfactory. The speech corpora are tagged automatically (Sphinx), but the output speech requires some manual finishing touches on the duration of the speech sounds adjacent to intra-sentence pauses. As Estonian is a word-central language the module of morphological analysis and disambiguation is very important for pronunciation. Unit selection and corpus-based synthesis was done (Mihkla et al. 2008) by using the Festival development system. The corpus-based synthetic voices have been derived from five donor voices:

- Riina (50 minutes of female speech)
- Tõnu (66 and 376 minutes of male speech)
- Liisi (116 and 434 minutes of female speech)
- Tõnis (71 minutes of male speech)
- Einar (89 minutes of male speech)

2.3. Formant synthesizer based on eSpeak

Estonian formant synthesis has also been realized on the eSpeak open- code development system for speech synthesis. The return to formant synthesis is mainly concerned with possible mobile applications. Besides compactness and multilinguality (over 40 languages) the eSpeak-et development version provides for speech synthesis supporting different platforms and using the markup languages SSML or HTML, and it can be used as a front-end to MBROLA diphone voices. The main shortcoming of formant synthesis is the segmental quality of the synthesized speech, which is responsible for the machine-like sound of the output. The speech is, nevertheless, clear and intelligible enough, while the available smart interfaces enable application of an eSpeak synthesizer in almost any operation system and its use in the voice applications based on different hardware.

2.3.1. HTS speech synthesis

In statistical parametric speech synthesis based on Hidden Markov Models (HMM), speech models are trained on moderate sized corpora. Although the segmental quality of the speech wave is lower than in unit selection, the HTS method has several advantages over the latter. First, the output speech is very fluent, with an almost ideally smooth speech melody. Second, real-time process management of the synthesis provides easy access to changing the speaking rate, voice pitch and timbre. As the HTS speech synthesis descends from the clustergen synthesis supported by Festival, statistical-parametric synthesis can make use of Festvox modules and the speech corpora meant for unit selection. Several Estonian-speaking HTS voices have been synthesized.

2.4. Comparison of the synthesis modules

Table 1 compares the characteristics of Estonian speech synthesizers. The most compact of the modules is obviously eSpeak-et, which is already accessible to mobile phone applications. The synthetic voices requiring unit selection and large speech corpora are, for the time being, evidently confined to server-based applications. The best segmental quality of the output speech has been achieved by means of unit selection. The quality of speech melody (speech prosody) and smoothness is the best in HTS synthesis. As for eSpeak its options of text processing are limited. Synthesis based on unit selection denies direct access to modification of individual parameters (speech rate, voice pitch, loudness and timbre). While variability in the speech wave often causes problems for speech recognition, too little variability in speech synthesis leads to monotonous and unnatural output speech (Tatham, Morton 2005). The most variable and most natural speech is no doubt achieved by the method of unit selection.

Table 1. Comparison of the modules of Estonian speech synthesis

Module	*Compactness*	*Segmental quality*	*Smoothness of output speech*	*Level of text processing*	*Level of process control*	*Speech variability*
MBROLA-et	moderate	high	moderate	high	high	moderate
Unit selection-et	low	high	moderate	high	low	high
eSpeak-et	high	low	moderate	moderate	high	low
HTS-et	moderate	moderate	high	high	high	moderate

3. Application and challenges of speech synthesis

3.1. ELTE audio

The ELTE audio system http://elte.eki.ee/ of e-texts and audio books is designed for the visually impaired people, enabling them to read the news, papers, magazines and books and listen to audio books over the Internet. The keywords of the web system are a largest possible usership and ease of implementation, so that the system could be used not only with home PCs but also, e. g. in libraries, at public internet access points and in public transport. The audio system has been made especially easy to handle for the blind. Selection of any news report, newspaper article or book, as well as browsing of newspapers and navigation in the audio library is enabled by using just the number keys and a couple of function keys. The software has options for browsing and retrieval in the server of information literature, for choosing the rate of reading and listening as well as the voice to be synthesized. The user can choose between different modules of Estonian TTS synthesis. Both unit selection and HTS synthesis offer a menu of several male and female voices. Access to the news, papers and audio magazines is free for all. The reading of e-books and listening to audio books requires a login; the passwords are available from the Estonian Library for the Blind to all its registered readers.

3.2. Voicing of subtitles

In Estonia dubbed films are less popular than films and telecasts aired in the original language with subtitles. However, there are many people with visual impairment who can see what is going on in the screen, but cannot see well enough to read the subtitle, as well as people who cannot read for other reasons (e. g. little children or people suffering from dyslexia). Digital television can provide an additional sound canal, where sound is generated from subtitle text (see Fig. 2). Separate canals for the telecast sound and the voiced subtitles enable the user to mix the sounds of the speech synthesizer and of the TV program (receiver mix). The service is accessible in many countries, e. g. Finland. Together with the Estonian Public Broadcast we have now entered a test period during which people with visual impairment pick the best synthetic voice, while a database of pronunciation of foreign names is compiled and the whole system is tuned in.

3.3. Automatic generation of audiobooks

Next year an amendment takes effect in Estonia that obligates publishers to submit, at the same time that a hard copy is published, the text files to the National Library. The law implements the principle of equal treatment, which means that the books should be equally accessible to the sighted as well as to the blind. In collaboration with the National Library we are now developing an automatic system for audiobook generation by means of a speech synthesizer. Like in subtitle voicing considerable attention is paid to the pronunciation of foreign names and the choice of the voice to be synthesized.

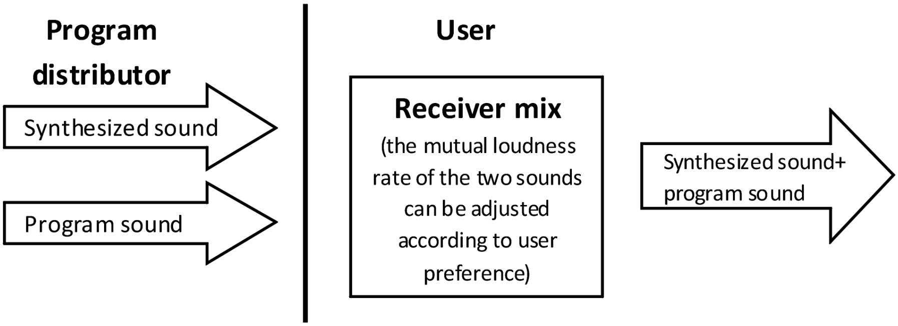
Program distributor
User
Synthesized sound
Program sound
Receiver mix
(the mutual loudness rate of the two sounds can be adjusted according to user preference)
Synthesized sound+ program sound

Figure 2. Receiver mix of voiced subtitles and the program sound by means of a digibox

3.4. Modelling speech prosody

One of the key concepts in speech technology is variability. Too little variability makes synthesized speech sound monotonous and unnatural. For the naturalness of the output speech we render as best possible the variation in the duration of speech sounds and pauses, the change of pitch throughout the phrase, the dynamics of speech intensity and the positions of pauses in the speech flow. In Estonian speech the temporal structure and rhythm have a special role as the phonologically significant tripartite opposition of the Estonian quantity degrees (Q1, Q2, Q3) is manifested on foot-level. In modelling quantity opposition one has to identify statistically significant features in speech temporal structure. A closer look at the phonetic parameters of Q1, Q2, Q3 in standard Estonian has revealed that specification of quantity opposition should rely on the durational relations between the vowels of stressed and unstressed syllables rather than on the durational relations between adjacent sounds (Kalvik et al. 2010). Also, the duration rate named first is a more stable parameter than pitch contours, which characterize and distinguish quantity degrees conditionally. A new approach to the higher pitch value, the position of which in the first vowel of the stressed syllable differentiates between the two long quantity degrees (Q2, Q3), suggests that comparison should be applied to the mean pitch values of different syllables (Mihkla et al. 2011). This facilitates the use of pitch (F0) as a phonetic parameter.

At word and phrase levels speech temporal structure raises various prosodic questions challenging speech synthesis. Estonian word stress is fixed on the first syllable, while the quantity opposition makes foot central for Estonian phonetics. It is still unclear whether Estonian speech rhythm is based on stress or on syllable. According to Asu, Nolan (2006) the rhythmic basis of Estonian speech might be the foot. If a sentence or phrase has a rhythmic structure of fixed feet, combined with quantity opposition, it is vital to know if and how the stress levels of the main stress syllables may change to fit the words into phrase.

One of the essential challenges facing the improvement of speech melody is to find out the relations of prosody with other language levels and extralinguistic features. If speech synthesis is based on nothing more than text phonetics, it cannot reach natural speech. Of prosodic relations, the change of the duration of the word-forming speech sounds depending on the part of speech and inflectional form of the word has been investigated. The experiments showed a few percent decrease of the output error if morphosyntactic and part-of-speech information had been added to the durational model (Mihkla 2007).

3.5. Emotional speech synthesis

Synthesis of emotional speech is one of the major challenges facing Estonian speech synthesis. At present, on the basis of the Estonian Emotional Speech Corpus http://peeter.eki.ee:5000/ (Altrov, Pajupuu 2012) the acoustic characteristics (pauses, speech rate, formants, intensity, pitch) of three basic emotions (anger, sadness and joy) and of neutral speech are being analysed. The aim is to ascertain what acoustic parameters are involved in speech emotionality and which of them are specific to this or that emotion, so that emotions could be modelled for speech synthesis. One of our principles is that speaker emotions can be identified sufficiently well from natural, i. e. non-acted speech and that non-acted speech is a prerequisite of natural speech synthesis (Iida et al. 2003). The emotions of the corpus sentences have been identified in perception tests, i. e. an emotion or neutrality has been regarded as identified if at least 51 % of the testers have agreed on it. The following four of the five acoustic parameters affecting speech emotion have been researched by now:

- pauses
- speech rate
- formants
- speech intensity

The analysis of pauses (Tamuri 2010) shows that for all three emotions concerned most of the pauses coincide with punctuation marks and there are more of breathing pauses than of non-breathing ones. According to the measurements of mean durations the longest pauses, for all three emotions, are those signalled of by a full stop and the shortest ones are the non-breathing pauses and the comma pauses. As for speech rate it was investigated if and how emotions may affect speech temporal structure. For this purpose the articulation rate of emotional utterances was measured in speech sounds per second and the results were compared with neutral speech data. Ranking the average speech rates from the most rapid to the slowest yields the following sequence: anger > joy > neutral > sadness (see Fig. 3).

In a formant analysis of emotional speech (Tamuri 2012) the first and second formants of the short stressed vowels *a*, *i* and *u* occurring in read emotional and neutral speech were measured. In pairwise analysis of the emotions (neutral included) it was found that statistically significant differences exist between the F1 means of the vowels *a* and *i*. Intensity was measured at the midpoint of each vowel both in emotional and neutral speech, paying attention to the means and range of intensity. According to the results intensity is a significant feature in distinguishing emotions. The rank order from the highest to the lowest intensity is as follows: neutral > anger > joy > sadness.

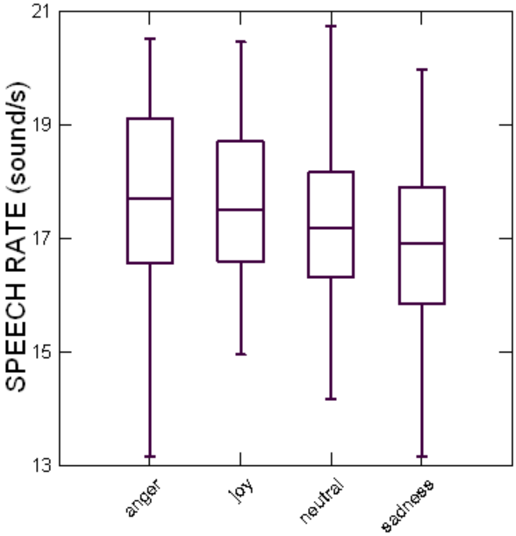

Figure 3. Correlation between speech rate and emotions

3.6. Speech synthesis interfaces

The uses of both speech synthesis and speech recognition could be enhanced if everybody interested had access not only to the modules of synthesis and recognition but also to smart interfaces (Speech Application Programming Interface). By means of such interfaces users could control the process of speech synthesis, to monitor text-to speech transformation, to follow text structure and vary the parameters (voice strength [loudness?], speech rate, voice pitch) of the synthetic voice in various voice applications. We are currently working on a development project aimed at producing prototypes of speech synthesis interfaces based on different platforms, enabling to control Estonian speech synthesis in different engines (eSpeak, Clustergen, unit selection, Mbrola, HTS), providing the widest possible coverage of the existing operation systems and hardware. For building such a SAPI (Speech Application Programming Interface) the existing speech synthesis modules (linguistic text processing, prosody generator and signal processing) are conformed to the information architecture of synthesis interfaces. Speech synthesis interfaces should guarantee that voice applications understand document structure, provide effective user navigation and support the more widespread mark-up languages (XML, SSML). At present near-perfect speech synthesis interfaces are available only with eSpeak supporting formant and, partly, diphone synthesis of Estonian. Interfaces for other modules are still being elaborated.

4. Conclusion and prospects

In the 21st century Estonian speech synthesis has been developed using the more widespread methods and freeware development systems (MBROLA, Festival, eSpeak, HTS). The efforts have mainly been focused on the development of language-specific modules, importing methods of signal processing. The applications have hitherto been developed mainly in view of the needs of the visually impaired (audio system for reading electronic texts, voicing of subtitles, creation of audiobooks). Nowadays that speech synthesis has become part of virtual reality its potentials could be increasingly used, e. g., in language learning, computer games or information retrieval. One of our following challenges is implementation of speech synthesis in web dictionaries, so that learners of Estonian as a foreign language could hear the pronunciation of words and their inflected forms as well as of sample sentences. Hitherto a wider use of speech synthesis has been hindered by the limited expressive power of synthetic speech. Synthesis of the basic emotions belongs to the main challenges facing Estonian speech synthesis. Analysis of the emotion-bound acoustic parameters (pauses, speech rate, formants, intensity and pitch) enable one to model emotions for speech synthesis. One of the essential challenges facing the improvement of speech melody is to find out the relations of prosody with other language levels as well as with extralinguistic features. Of prosodic relations, the change of the duration of the word-forming speech sounds depending on the part of speech and inflectional form of the word has been investigated. The experiments showed a few percent decrease of the output error if morphosyntactic and part-of-speech information had been added to the durational model. More widespread use of speech synthesis could also be enhanced if everybody interested could implement any existing module of Estonian speech synthesis in their own application programs. For that purpose smart speech synthesis interfaces should be created for all of the modules, with possibly widest coverage of the existing operation systems and hardware.

5. Acknowledgements

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

References

1. *Altrov R., Pajupuu H.* (2012), Estonian Emotional Speech Corpus: Content and options. G. Diani, J. Bamford, S. Cavalieri (Eds.). Variation and Change in Spoken and Written Discourse : Perspectives from Corpus Linguistics. Amsterdam : John Benjamins, [forthcoming].
2. *Asu E. L., Nolan F.* (2006), Estonian and English rhythm: a two-dimensional quantification based on syllables and feet. R. Hoffmann, H. Mixdorff (Eds.). Speech Prosody. Dresden: TUDpress, 249–252.
3. *Iida A., Campbell N., Higuchi F., Yasumura M.* (2003), A corpus-based speech synthesis system with emotion. Speech Communication 40, 161–187.
4. *Kalvik M.-L., Mihkla M.* (2010), Modelling the Temporal Structure of Estonian Speech. In: Human Language Technologies. The Baltic Perspective : Proceedings of the Fourth International Conference, Baltic HLT : Riga, Latvia, October 7–8, 2010. (Eds.) Skadina I., Vasiljevs A . Amsterdam: IOS Press, Frontiers of Artifical Intelligence and Applications, 219, 53–60.
5. *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. (Eds.) Narin'yani, A. S. Tarusa, 351–353.
6. *Mihkla, Meelis (2007). Modelling speech temporal structure for Estonian text-to-speech synthesis: feature selection. Trames : Journal of the Humanities and Social Sciences, 11(3), 284–298.*
7. *Mihkla M., Kalvik M.-L.* (2011), Significant features of Estonian word prosody. In: Proceedings of the 17th International Congress of Phonetic Sciences: The 17th International Congress of Phonetic Sciences (ICPhS XVII), Hong Kong, China, August 17–21, 2011. (Eds.) Wai-Sum Lee & Eric Zee. Hong Kong: City University of Hong Kong, 1378–1381.
8. *Piits L., Mihkla M., Nurk T., Kiissel I.* (2007), Designing a speech corpus for Estonian unit selection synthesis. In: Nodalida 2007 Proceedings: The 16th Nordic Conference of Computational Linguistics, 367–371.
9. *Remmel M., Tago, T.* (1984), Towards increasing the commercial success of speech synthesizers. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84.: 91–93.
10. *Taylor P.* (2009), Text-to-Speech Synthesis. Cambridge University Press.
11. *Tatham M., Morton K.* (2005), Developments in Speech Synthesis. John Wiley & Sons, Ltd.
12. *Tamuri K.* (2010), Kas pausid kannavad emotsiooni? Eesti Rakenduslingvistika Ühingu Aastaraamat, 6, 297–306.
13. *Tamuri K.* (2012), Kas formandid peegeldavad emotsioone? Eesti Rakenduslingvistika Ühingu Aastaraamat, 8, 231–243.