Human Language Technologies – The Baltic Perspective
K. Muischnek and K. Müürisep (Eds.)

doi:10.3233/978-1-61499-912-6-79

Self-Reading Texts and Books

Meelis MIHKLA[1], Indrek HEIN and Indrek KIISSEL
Institute of the Estonian Language

Abstract. The rise of e-books, the cumulative digitisation of written library materials and the advancement of speech technology have reached a stage enabling library services and e-books to be read out loud to customers in synthetic speech and paper books (either published or still in print) to be delivered in the audio form. The user environment of the digital archive Digar of the Estonian National Library includes a special reading machine capable of producing an audio version of electronic texts in Estonian (books, magazines etc). The application of Elisa Raamat provides access to more than 2500 Estonian e-books, which can not only be read visually from the screen of a smartphone or tablet but also listened to. The speech server of the Institute of the Estonian Language offers, as a public service, the text-to-speech system Vox populi, inviting people to have an audio version synthesized from any text of interest, being prepared to convert any uploaded text (an article, paper, subtitle file, e-book etc) into an audio file. The present study is focused not only on the description of the systems but also on various issues of text processing and pronunciation as well as on the reflection of text structure in synthetic speech. The quality of self-reading largely depends on how adequately the input abbreviations, numbers and other non-letter sequences are converted into words in correct morphological form and how closely the output pronunciation of foreign names matches that of the source language. In the article we will also discuss a special module for text pre-processing, which helps in the case of more complex text structures and character sequences (e.g. geographic coordinates, sports results, numeral inflection). In addition, book reading requires an as accurate as possible rendering of text structure. The study also analyses audio books to capture the essence of human prosodic phrasing as well as different pauses and the marking of reported speech when talking.

Keywords. Audio books, speech synthesis, phonetic database of foreign names, direct speech, phonetic phrasing

1. Introduction

Estonians are well known for their readership and good reading skills [1]. Indeed, we read a lot of books and periodicals. And yet there are lots of people in our society for whom reading presents a challenge, one way or another. The primary target groups of our projects were people with a visual or reading challenge (the registered number of the former makes up 0.5% of the Estonian population, while the latter are estimated at 10%). In addition, self-reading interfaces could be of use to the elderly with deteriorating visual acuity and to children with insufficient reading skills. Audio access to books is also welcomed by anyone in certain situations (driving, travelling, working out, dog walking) where one cannot read paper books or watch the screen of a smart device. We have found it a pleasant surprise that the option of self-reading books is

[1] Corresponding Author: Meelis Mihkla, Institute of the Estonian Language, Roosikrantsi 6, Tallinn 10119, Estonia; E-mail: meelis.mihkla@eki.ee.

used frequently by every third user (ca 6500 people) of the paid service of Elisa Raamat. Consequently, Estonian speech synthesis has ceased to be a mere niche product for the handicapped, but has become part of the everyday life of many ordinary people.

The United Nations Convention on the Rights of Persons with Disabilities [2], Article 30, paragraph 1c requires, inter alia, that the handicapped should have a guaranteed access to library services and print media. Specifically, online library services for the blind, visually impaired and dyslexics require access to a special reading machine from the library archives website. The Legal Deposit Copy Act [3] recently adopted by the Riigikogu requires publishers, since January 1, 2017, to immediately submit the output-ready file or an equivalent copy file of practically every new publication released in the Republic of Estonia to the National Library, using its online submitting system. The availability of a system for generating electronic book files as well as audio books enables an automatic conversion of paper books into audio form. This option complies with the internationally recognized principle of equal treatment. The Equal Treatment Act [4] in force in the Republic of Estonia provides, in principle, that new books and other publications should be simultaneously available in audio form to ordinary people as well as to those linguistically challenged. Automatic synthesising of audio books requires electronic text files of books and a synthesiser of the language in question. By to-day, a very large portion of the Estonian-language material held at the Estonian National Library has already been digitised, having thus become accessible to online use and processing. Even the Estonian blackletter publications, which are problematic to most modern readers, are subjected to digitisation. The enforcement of the Legal Deposit Copy Act guarantees that practically all printed word published in Estonia is preserved, in electronic form, at the digital archives of the Estonian National Library and thus, in principle, is available for use by the linguistically challenged.

Of course, it is premature to expect that synthetic speech could rival human speech in naturalness, pleasantness or expressiveness, either now or in the near future. Actors and professional readers will go on producing audio books without reason to fear direct competition from their synthesized equivalents. The problem is rather in means and resources. In Estonia, the annual output of publications includes about 3500 Estonian-language books. The Estonian Library for the Blind employs actors to record about 160 audio books a year, which is less than five percent of the annual output of books. The availability of automatic systems of text-to-speech synthesis and audio book generation, however, create an opportunity to make all printed and electronic publications, or information therein, accessible in audio form (read out by a synthetic voice) simultaneously with their publication in print, while the most interesting texts (e.g. those with the longest queue of audio requests at the speech server of the National Library) could still be voiced by actors or professional readers.

2. The TTS system *Vox populi* for books and other texts

In principle, the text-to-speech system *Vox populi*, the self-reading app Iselugej*a* of Elisa Raamat and the TTS reading machine of Digar are all text-to-speech converters of a basically similar function and structure. As *Vox populi* has the broadest uses and functions of the three, the article concentrates in more detail on this public service, which converts input texts (in txt- or html-format), subtitle files (with stl-extension) or

e-books (in epub-format) into audio files and epub3 audio books. *Vox populi* consists of two components – an editing interface and a synthesising interface (see Fig. 1). The editing interface analyses the input text and makes up a list of the assumed foreign names and unknown character sequences to be matched with the pronunciation database. The pronunciation database contains over 100, 000 entries. The on-line environment of the pronunciation of foreign names enables addition of new pronunciations as well as checking the accuracy of pronunciation. The editing interface includes a special module for text pre-processing, which helps in the case of complex text structures and character sequences such as, for example, geographic coordinates, sports results, or numeral inflection. The synthesising interface enables choosing between different synthetic voices and modifies the speech rate for the voice chosen [5].

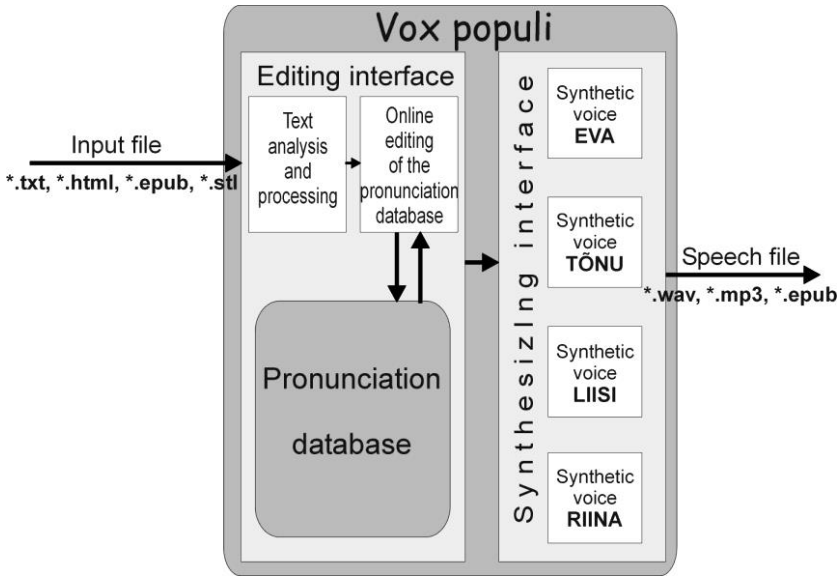

Figure 1. The TTS system *Vox populi,* with editing and synthesising interfaces, to produce audio versions of books and other texts.

There are six speech rates to choose from: normal, faster, very fast, terribly fast, slow, and very slow. Studies [6] and [7] have proven that some blind people can understand speech that is twice or even faster than normal and is perceived by the sighted as incomprehensible (´terribly fast´). The presence of such talented listeners among Estonian blind people has been proved by our personal observations [8].

3. The text pre-processing module and the database of foreign name pronunciation

First, *Vox populi* subjects the input text to Morphological Analysis[2], from whose output non-character clusters (numbers, geographic coordinates, special symbols etc) as well as abbreviations, unknown character sequences and foreign names have to be identified. Speech synthesis requires that abbreviations and non-character sequences be converted to readable words. True, the synthesiser is also able to convert digits and special symbols to readable text, but this is mechanical conversion neglecting context and the specificity of book texts. Reading books to someone and correct interpretation of texts, however, takes a considerably more detailed analysis and processing of digits, special symbols, abbreviations and foreign names.

The most complicated digit sequences to read are geographic coordinates (e.g., the geographic coordinates of Tallinn Town Hall Square are 59° 26′ 14″ N, 24° 44′ 43″ E –> *viiskümmend üheksa kraadi kakskümmend kuus minutit neliteist sekundit põhjalaiust ja kakskümmend neli kraadi nelikümmend neli minutit nelikümmend kolm sekundit idapikkust*). Another frequent problem is accurate presentation of sports results (e.g. marathon results can be presented in several ways: 2:24.04 or 2.24.04 or 2.24:04, and in all three cases the system should say, -> *kaks tundi kakskümmend neli minutit ja neli sekundit.* Or the relay race -- *4 x 100 m teatejooks* --, where the character *x* is pronounced *korda* ´times´ –> *neli korda saja meetri teatejooks*). The correct case form of numerals can be determined by the help of the case forms of adjacent words (e.g. *Ehitis valmis 2011. aastal* ´The building was finished in 2011´. The output of the Morphological Analysis module is: *aasta+l //_S_ sg ad, //,* consequently the ordinal numeral *2011* is in the adessive case –> *kahe tuhande üheteistkümnendal*). Sometimes we find morphological markers attached to digit sequences, either hyphenated or not (134-ne –> *saja kolmekümne neljane*, 100s –> *sajas*). Some words (especially adverbials) govern the morphological form of the numeral expressing the preceding or following digit or number: If the digit or number is preceded by *üle* ´over´ or followed by *paiku* ´about´ or *võrra* ´by´, the digit or number should be read in the genitive case (e.g. *üle 5 m –> üle viie meetri* ´over 5 metres´, *kohtume 3 paiku –> kohtume kolme paiku* ´let´s meet about 3 o´clock´, *nihkusime 2 koha võrra edasi –> nihkusime kahe koha võrra edasi* ´we moved to the next but one seat´). If the digit or number is preceded by *kell* ´o´clock´, *neist* ´of them´ or *joobes (joove)* ´intoxicated (BAC)´, the number takes the nominative case (*kell 23 –> kell kakskümmend kolm* ´at eleven PM´; *neist 3 oli katki –> neist kolm oli katki* ´three of them were broken´; *tuvastati joove 1,2 ‰ –> tuvastati joove üks koma kaks promilli* ´the BAC detected was 1.2 ‰´). Particular attention should be paid to the declension of decimal fractions (numbers with a decimal point), which is rather complicated and not always unambiguous (e.g. in *2,3 % võrra –> kahe koma kolme protsendi võrra* ´by two point three percent´ both numerals take the genitive, which, however, does not hold in the interval between 1 and 0 (excl.), e.g. 0,3 ´3.0´ –> null koma kolme (*null* in the nominative), not **nulli koma kolme*).

During text analysis the foreign names and unknown character sequences will be entered in a special list, where the customers can add pronunciations and check their sound. The default option of the editor interface is *Nagu häädus ütleb* ´see

[2] Free open source software, can also be used and modified as part of commercial applications (https://github.com/Filosoft/vabamorf)

transcription´ referring the foreign name or unknown character sequence to the transcription available in the pronunciation field of the database (e.g. Tracy -> trassii, Fere -> feer). Other possible options are: *Sõna hääldada nagu on ´*pronounce like an Estonian word´ (e.g. Rudolf -> Rudolf, Hering -> hering) instructs to follow the Estonian rules of pronunciation, *Eesti lühend, tähthaaval ´*Estonian abbreviation, spell by letter names´ (e.g. SA -> es aa, not *sihtasutus*; as there can be several different *SA*-abbreviations and even the pronoun *sa ´*you (Sg)´ can sometimes, for the sake of emphasis, be written in capitals there is no point in sticking to a particular equivalent of the *SA* abbreviation); *Lühend, hääldada nagu on ´*abbreviation, pronounce like an Estonian word´ (e.g. AIDS –> aids) or *Inglise lühend, tähthaaval ´*English abbreviation, spell by letter names´ (e.g. IBM -> ai bii emm). Simple spelling mistakes can be corrected by the interface signalling *Kirjaveaga, õige kuju hääldusväljal ´*spelling mistake, see pronunciation field´ (e.g. Mancester = Manchester -> mantšester, Intagram = Instagram). To foreign names, case endings are often added following an apostrophe (e.g. Bastille'sse -> bast'iisse (Illative), Pierre Rouge'i -> piäär ruuži (Genitive, Illative)), but the text processor may find it confusing to have to work with 4 or 5 different apostrophes. Special attention needs to be paid to the pronunciation and declination of compound foreign names (e.g. Grande-Rue's -> grand rüüs, krahv de la-Fere'is -> döla'feeris, proua de Bois-Tracy'iga -> döbu'aa trass'iiga). The foreign name pronunciation database, presently containing the pronunciations of over 100 000 foreign names, abbreviations and character sequences is constantly accumulating new items.

4. Prosodic phrasing and direct speech

For the sake of readability printed text usually follows a certain structure. A text can be divided into parts, chapters, paragraphs, bullet points etc. In addition, the text may contain highlighted passages of direct speech, as well as footnotes, citations and references. In an ideal case, the synthesiser should also be able to convey this text structure as adequately as possible.

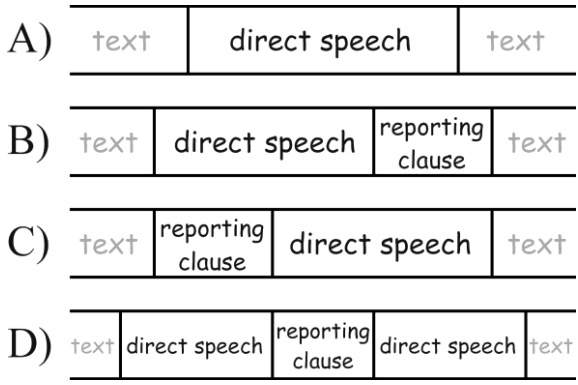

Figure 2. Direct speech in different text structures.

In a previous study we found, based on listener results of perception tests, the optimal acoustic parameter values for highlighting the title and for marking the boundary between direct speech and reporting clause [9]. In the present work we explore audio books to find out how the text is divided into phrases and how direct speech is presented in real read speech. Four audio books, [10], [11], [12] and [13] were analysed. Each book is represented by a passage of 12 – 29 minutes of speech.

The audio marking of direct speech by means of changing the fundamental frequency (F0) level was investigated in four possible text structures (see Fig. 2). In the first case (A) there is no reporting clause and the passage of direct speech occurs between narrative passages, while in (B) the reporting clause follows direct speech, in (C) it precedes direct speech, and in (D) the reporting clause occurs between two passages of direct speech. Measurements for analysis included the F0 baseline level in passages of direct speech, in reporting clauses and in adjacent sentences of the preceding and following passages. The F0 baseline differences were calculated in semitones.

Table 1. Comparison of the mean F0 baseline values with those of adjacent text segments in different text structures.

Direct speech structure type	F0 baseline comparisons	F0 baseline differences (in semitones)	p-value
(A)	$F0_{DC} - F0_{pt}$	0.57	0.632
	$F0_{DC} - F0_{ft}$	-0.09	0.935
(B)	$F0_{DC} - F0_{pt}$	1.25	0.086
	$F0_{DC} - F0_{RC}$	**3.15**	**0.003**
(C)	$F0_{DC} - F0_{RC}$	1.21	0.092
	$F0_{DC} - F0_{ft}$	0.31	0.747
(D)	**$F0_{DC1} - F0_{RC}$**	**4.14**	**0.003**
	$F0_{DC2} - F0_{RC}$	**3.96**	**0.005**

DC – direct speech, RC – reporting clause, pt – preceding text, ft – following text

Table 1 presents the mean F0 baseline differences between direct speech and adjacent passages. It reveals that if there is no reporting clause (type A), those differences are not significant statistically (p=0.632; p=0.935). If direct speech is followed by a reporting clause (type B), its mean F0 baseline is 3.15 semitones higher than that of the following reporting clause (p=0.003), but not significantly higher than the mean F0 baseline of the text sentence preceding the direct speech passage (p=0.086). If the reporting sentence precedes direct speech (type C), the F0 differences between the direct speech and its adjacent passages are, surprisingly, not significant statistically (p=0.092; p=0.747). However, the mean F0 baseline of direct speech is significantly higher -- 4.14 semitones (p=0.003) and 3.96 semitones (p=0.005) -- than that of the reporting clause if the latter occurs between two passages of direct speech (type D). Interestingly, perception tests assessing the acoustic marking of direct speech scored highest for a 2.5 semitone difference between direct speech and the reporting clause [9], which is relatively close to the mean values computed from natural speech (2.5 semitones vs 3.15-4.14 semitones).

Table 2. Mean durations and standard deviations of pauses at the end of phrases, sentences and paragraphs.

Actor/Speaker	Phrase end pauses		Sentence end pauses		Paragraph end pauses	
	Mean value (in ms)	Standard deviation	Mean value (in ms)	Standard deviation	Mean value (in ms)	Standard deviation
VP	513	297	1238	414	1739	464
HK	461	247	1142	566	1954	491
KK	575	222	1248	176	1690	438
AM	508	183	882	318	1669	524
Total	503	239	1139	443	1741	482

The same audio books were used to analyse intra-sentence prosodic phrasing as well as pausing at sentence and paragraph end. Table 2 displays the mean durations and standard deviations of phrase-, sentence- and paragraph-end pauses, averaged for each informant as well as for the total material. Relative variation turned out to be highest for phrase-end pauses (see Figure 3).

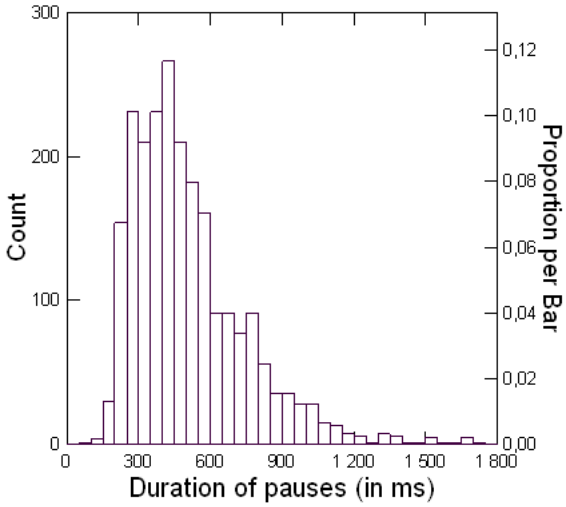

Figure 3. Histogram of the duration of phrase-end pauses.

In general, the human readers (actors in particular) seem to have treated the prosody of the sentences rather freely, due to which the phrase end pauses do not often coincide with punctuation or conjunctions. The text is interpreted creatively, emphasising this or that word or phrase. Compared with an earlier pause study by the same author [14] the average pauses measured from audio prose for the present study are 59 – 70 % longer. This is particularly noticeable in intra-sentence pauses, which are nearly twice as long as newsreading pauses.

5. Conclusion and next steps

Estonian speech synthesizers and audio readers diversify the presentations of printed word and listening to books in synthetic speech enhances people´s options for sharing in written information. Even though reading out books in synthetic speech need not be quite innocent insofar as the accent and emphasis errors may affect our own language usage, there is no stopping technological progress. Speech synthesis and TTS systems are subject to constant development and improvement concerning the quality of the synthetic speech, the methods of analysis and synthesis, and system functionality in order to offer the user a better and more reliable service.

What is certain is that the existing solutions provide an additional option for a great number of linguistically challenged people (the visually impaired, dyslectics, elderly people and small children) to share in the digitized part of library collections as well as of new publications or ePubs ahead of print. The app Elisa Raamat has moved many Estonian books of fiction into modern smart devices, which will hopefully increase the numbers of young readers and listeners. An analogous tendency applies to schoolbooks and workbooks, which are also moving over from satchels to computers as ebooks, which is another app to be supplied with an audio option[3].

The present study has analysed audio books, addressing prosodic phrasing, duration of different pauses, and the acoustic marking of direct speech. In further works it is planned to predict phrase boundaries from written text. Another way to improve the quality of output speech would be automatic detection of prosodic prominence from written text. In addition, the text preprocessing modules need to be improved. Attention should also be paid to the output format of audio books. In the existing text-to-speech systems the output speech is directed either directly to the audio output of the user´s listening device or to some audio files (in mp3- or wav-format) that have no direct connection either with the printed page or the screen page of an e-book. A new multimedia format for ebooks, ePub3, is now being developed to create an integrated media presentation by means of a logical connection established between the relevant texts, images, synchronic audio and video files. The application of such a standard format allows to safely say that a book can not only be read and listened to but also watched in the form of moving images.

Acknowledgements

Research and development in speech synthesis and TTS systems for automatic conversion of books and texts to audio format has received financial support from The National Programme for Estonian Language Technology, the institutional research theme IUT35-1 and the European Regional Development Fund (CEES, Centre of Excellence in Estonian Studies).

[3] see https://www.opiq.ee/

References

[1] European Commission, *Final report of EU high level group of experts on literacy*, Luxembourg: Publications Office of the European Union, 2012.
[2] ÜRO puuetega inimeste õiguste konventsioon, *Riigi Teataja II* **6** (2012).
[3] Säilituseksemplari seadus, *Riigi Teataja I* **1** (2016).
[4] Võrdse kohtlemise seadus, *Riigi Teataja I* **22** (2012).
[5] M. Mihkla, I. Hein, I. Kiissel, A. Räpp, R. Sirts, T. Valdna, A System of Spoken Subtitles for Estonian Television, *Frontiers in Artificial Intelligence and Applications* **268** (2014), 19–26.
[6] A. Moos, I. Hertrich, S. Dietrich, J. Trouvain, H. Ackermann, Perception of Ultra-Fast Speech by a Blind Listener – Does He Use His Visual System?, *Proceedings of the 8th Seminar on Speech Production, ISSP* (2008), 297-300.
[7] I. Hertrich, S. Dietrich, H. Ackermann, Cross-modal interactions during perception of audiovisual speech and nonspeech signals: an MRI study, *Journal of Cognitive Neuroscience* **23** (2011), 221-237.
[8] M. Mihkla, I. Hein, I. Kiissel, M. Orusaar, A. Räpp, Kõnetempo eelistused ja audiosüsteem nägemispuudega inimestele / Preferences of speech rate and an audio system for the visually impaired, *Keel ja Kirjandus* **5** (2011), 334–342,.
[9] M. Mihkla, I. Hein, A. Hiiepuu, I. Kiissel, R. Ruusalepp, U. Sinisalu, Raamat sünnib kuulata / Books for listening, *Keel ja Kirjandus* **2** (2017), 114–129.
[10] A. Jürgen, *Sinine Lind / Blauvogel Wahlsohn der Irokesen*, Audiobook, by artist Valli Pärn, sound operator Mart Vaaks, Estonian Library for the Blind, 2010.
[11] K. Hamsun, *Aga elu kestab / Men livet lever*, Audiobook, by artist Hans Kaldoja, sound operator Mart Vaaks, Estonian Library for the Blind, 2012.
[12] E. Bornhöhe, *Vürst Gabriel, ehk, Pirita kloostri viimsed päevad*, Audiobook, by artist Karol Kuntsel, sound operator Mart Vaaks, Estonian Library for the Blind, 2008.
[13] C. Erickson, *Šoti kuninganna Mary memuaarid / The memoirs of Mary Queen of Scots*, Audiobook, by artist Anne Margiste, sound operator Mart Vaaks, Estonian Library for the Blind, 2011.
[14] M. Mihkla, Pausid kõnes / Pauses in speech, *Keel ja Kirjandus* **4** (2006), 286–295.