Estonian isolated-word text-to-speech synthesiser

Indrek Kiissel
ikiissel@gmail.com

Liisi Piits
Liisi.Piits@eki.ee

Heete Sahkai
Heete.Sahkai@eki.ee

Indrek Hein
Indrek.Hein@eki.ee

Liis Ermus
Liis.Ermus@eki.ee

Meelis Mihkla
Meelis.Mihkla@eki.ee

Institute of the Estonian Language, Tallinn, Estonia

Abstract

This paper presents the development and evaluation of an Estonian isolated-word text-to-speech (TTS) synthesiser. Unlike conventional TTS systems that convert continuous text into speech, this system focuses on the synthesis of isolated words, which is crucial for applications such as pronunciation training, speech therapy, and (learners') dictionaries. The system addresses two key challenges: generating natural prosody for isolated words, and context-free disambiguation of homographs.

1 Introduction

Text-to-speech synthesis (TTS) is typically used to convert texts and sentences into speech. However, there are many applications that require the speech synthesis of isolated words: pronunciation training applications, speech and language therapy applications, (learners') dictionaries, etc. Such applications additionally require a careful and correct pronunciation of the synthesised words. To achieve this, the TTS system must fulfill two additional requirements beyond the general requirements for TTS systems. First, the training data must contain a sufficient amount of short utterances in order for the system to be able to generate isolated words with a natural utterance prosody. Second, the system must allow for a context-free disambiguation of input words that have phonologically different homographs. While the first requirement is unproblematic, the second requirement is a considerable challenge for a language like Estonian. Estonian possesses a large number of homographs that are mainly due to the absence of orthographic marking for two phonological features of Estonian: palatalisation and, in certain cases, third quantity (overlong length degree). This gives rise to two main types of homograph pairs: homographs differing in palatalisation, and homographs differing in second quantity (Q2) vs. third quantity (Q3). Palatalisation in Estonian is, on the one hand, a coarticulatory phenomenon, meaning that all alveolar consonants /t, s, n, l/ preceding /i/ or /j/ at the boundary of the primary stressed syllable and the following syllable become palatalised. On the other hand, it is also a phonological phenomenon that distinguishes meaning (Metslang et al., 2023). The distinction between second and third quantity results from a difference in the prosodic structure of long stressed syllables, which can occur either in a disyllabic (Q2) or monosyllabic (Q3) foot (Metslang et al., 2023). Both palatalisation and quantity distinctions can be challenging for learners of Estonian as a second language and thus require attention in language pedagogy applications (Malmi et al., 2022b; Meister and Meister, 2014).

The homograph pairs differing in palatalisation are always (inflectional forms of) different lemmas whereas quantity distinguishes both between homographic lemmas and inflectional forms of the same lemma. For example, the orthographic form *tulp* represents both /tulpː/ 'signpost.NOM.SG' and /tulʲpː/ 'tulip.NOM.SG', and *maitse* represents both /mɑitse/ 'taste.NOM.SG' and /mɑitːse/ 'taste.GEN.SG' or 'taste.IMP.2SG'. In addition, numerous words have pronunciation variants differing only in quantity or palatalisation. The Estonian Combined Dictionary (CombiDic) (Langemets et al., 2023) contains altogether 756 homographs and pronunciation variants differing in palatalisation, and 22,618 homographs and variants differing in quantity (excluding compounds). While the incorrect pronunciation of these homographs does not necessarily hinder comprehension in context, it does so without context and is particularly problematic in pedagogical applications.

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 302–306

A TTS system that is able to generate isolated words with a correct pronunciation must thus include a means for disambiguating homographs. Current supervised Estonian TTS systems include morphological parsing and disambiguation as part of their pre-processing pipeline. The standard morphological parser and disambiguator currently used in the Estonian TTS systems is Vabamorf[1] (Kaalep and Vaino, 2001). In addition to part-of-speech and inflectional categories the parser annotates compound boundaries and the following pronunciation features: third quantity, irregular stress, and palatalisation. Morphological parsing is followed by disambiguation; however, disambiguation is based on the probability of tag sequences within sentences and thus cannot be applied to isolated input words. As a result, the probability that an existing supervised TTS system generates the desired member of a homograph pair is at chance level. Likewise, the disambiguation of homographic input words is infeasible in unsupervised TTS systems, which may produce palatalisation, quantity, stress and compound identification errors also in words without homographs. In order to solve this problem, we developed a dedicated Estonian TTS system for generating isolated words with a correct pronunciation. Section 2 describes the development and the features of the system (training data, pre-processing, TTS technique, and user interface), Section 3 evaluates the performance of the system in terms of the pronunciation accuracy of homographic minimal pairs differing in palatalisation or quantity, Section 4 describes the planned and potential use cases of the system, and Section 5 presents the conclusion and future steps.

2 Development and features of the Estonian isolated-word TTS system

2.1 Training Data

The training data consisted of human-recorded sound files of isolated words and the corresponding text files. The sound files had been recorded for language pedagogy purposes by a female voice talent in a sound studio in order to exemplify the pronunciation of a subset of the headwords of the CombiDic (the basic vocabulary). The dataset consisted of a total of 31,215 words (10 h 36 min) with a good coverage of Estonian sounds and sound combinations and a high phonetic quality[2]. The materials thus provided appropriate training data for ensuring a natural production of isolated words as utterances and a good phonetic coverage and quality suitable for pedagogical and speech therapeutic applications. The text versions of the words were drawn from the database of the CombiDic along with diacritics for third quantity, irregular stress, palatalisation, and compound boundaries. The annotation principles are based on Viks (1992)[3] and are standardly used in Estonian dictionaries and parsers, including Vabamorf.

2.2 Pre-processing

The pre-processing did not include the standard stage of parsing as the input words were already annotated for the relevant features normally assigned by the parser. Otherwise, the standard pre-processing steps and grapheme-to-phoneme conversion used in Estonian TTS were applied (Mihkla et al., 2000).

2.3 TTS technique

We used the Merlin TTS toolkit developed by the Centre for Speech Technology Research (CSTR) at the University of Edinburgh[4] (Wu et al., 2016). It is designed for building deep neural network models for statistical parametric speech synthesis. Merlin TTS was considered a suitable technique as it requires a relatively small amount of training data and allows good control. The model was developed specially for isolated word synthesis[5] (Kiissel, 2024).

2.4 User interface

The synthesiser is available online via https://elo.eki.ee/yksiksona/ (see Figure 1). The user must enter the word to be synthesised along with the appropriate diacritics for third quantity, palatalisation, irregular lexical stress and compound boundaries to obtain the desired pronunciation. The interface provides instructions for inserting the diacritics. To help users insert the necessary diacritics the web page will additionally include a Vabamorf interface for automatically annotating input words with morphological tags, compound boundaries, and pronunciation marks.

[1] https://github.com/Filosoft/vabamorf/tree/master

[2] The corpus is available at https://koneveeb.ee/korpused/#eva_yksiksonad (eva_yksiksõnad_1, eva_yksiksõnad_2).

[3] see also https://eki.ee/teatmik/haaldusmargid-uhendsonastikus-us/

[4] https://github.com/CSTR-Edinburgh/merlin and https://www.cstr.ed.ac.uk/projects/merlin/

[5] https://github.com/ikiissel/mrln_et_iw

Users can download the synthesised pronunciations as WAV files.

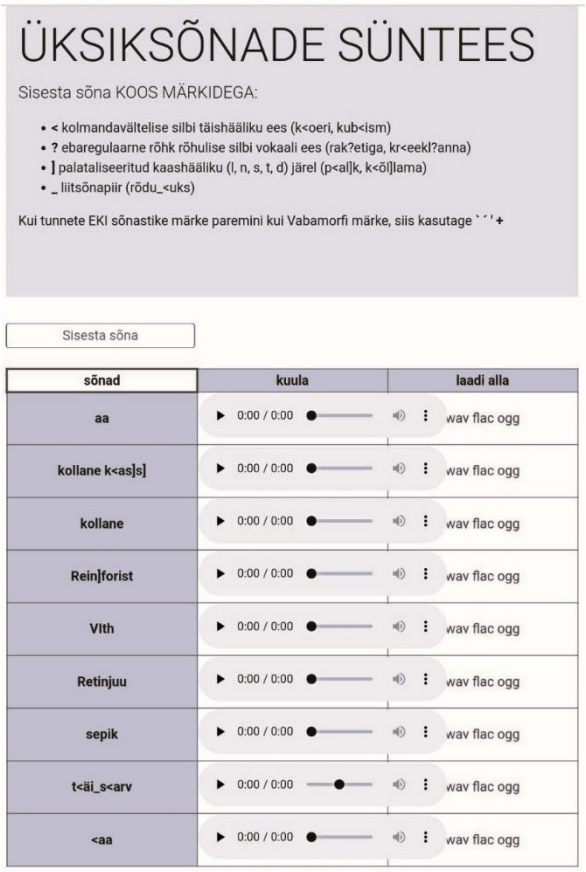

Figure **1**: User interface of the synthesizer.

3 Evaluation

3.1 Materials, evaluators, procedure

We conducted a perception test to evaluate the performance of the TTS system in terms of pronunciation accuracy. We used 16 pairs of homographs that differ in palatalisation and 16 pairs of homographs that differ in quantity. Two types of monosyllabic word pairs were included for the evaluation of palatalisation: words ending with a long consonant like *kott*, *konn*, *tall*, and words with a consonant cluster like *palk*, *mulk*, *sulg*. The homographs distinguished by quantity were selected to include words with different syllable structures: (C)VCCV, e.g., *paksu, kommi, arve*; CVVCCV, e.g. *maitse*; CVVV, e.g., *saia*; CVVCV, e.g. *heina*; CV̄CV, e.g., *hoone.*

All the items were synthesised using the diacritics corresponding to the two pronunciations, e.g., "p<al[k" for /palʲkː/ and "p<alk" for /palkː/, and "kommi" for /kommi/ and "k<ommi" for /komːmi/.

The perception test was carried out online in the LimeSurvey[6] environment. The task of the evaluators was to listen to each item and to answer one of the following questions, depending on the case: Is this word palatalised or not? Is this word in second or third quantity? There were in total 32 cases where the evaluators had to determine whether the word they heard has palatalisation or not, and 32 cases where they had to decide whether the word was in the second or third quantity[7].

The evaluators were eight linguistics and phonetics experts who had previous experience in identifying both palatalisation and quantity.

3.2 Evaluation results

Palatalisation. Out of 32 homographs, 26 were correctly recognised by all the experts (100%). For the words /tulʲpː/, /kotʲː/ and /patʲːs/ the intended pronunciation was recognised by 88% of the experts, and for the words /jutʲː/, /nutʲː/, /mütːs/ by 75% of the experts. It appears that problems mainly arise with words involving /t/ and /tʲ/ (except for /tulʲpː/). Given that all the test items were correctly recognised by a majority of the evaluators, the performance of the synthesiser can be considered very good. Occasional failures to perceive palatalisation were to be expected as palatalisation in Estonian has been found to be variable, weak, and gradient, and it has been noted that, especially in connected speech, experts' opinions on the identification of palatalisation may not always coincide (Kalvik and Piits, 2019).

Quantity. The intended quantity of each test word was recognised by almost 100% of the evaluators. Only in the case of the word /mɑitse/ 'taste.NOM.SG' did one out of the eight experts fail to recognise that it was a Q2 form. For the remaining 31 word forms, all the experts recognised the intended quantity.

In summary, the performance of the isolated word synthesiser in terms of the phonetic accuracy of homographic words is very good, whereas the probability of obtaining a desired pronunciation variant with other Estonian TTS

[6] https://www.limesurvey.org/

[7] The materials and evaluations are available at https://doi.org/10.6084/m9.figshare.27275964

systems is only 50% due to the absence of disambiguation.

4 Use cases

The isolated-word TTS synthesiser allows the user to generate correctly pronounced isolated words and multi-word units by manually specifying the features of quantity, palatalisation, lexical stress and compound structure. The synthesiser generates isolated words with an appropriate utterance prosody and high phonetic quality, being thus suitable for language pedagogical and speech therapeutic purposes. Below, we describe three planned or potential use cases of the isolated-word synthesiser.

Generation of pronunciation examples for dictionaries. CombiDic currently uses TTS to generate the audio for example sentences. For headwords, the dictionary currently includes human-recorded pronunciation examples (used as the training data of the isolated-word synthesiser, see Section 2.1). However, pronunciation files are available only for the basic vocabulary, and only for three or four inflectional forms of inflecting words, depending on part-of-speech. The first application of the isolated-word synthesiser will therefore be the generation of pronunciation files for all the headwords and for all the inflectional forms in the CombiDic. In addition, pronunciation files are essential for learners' dictionaries, e.g., the Estonian Picture Dictionary[8].

Pronunciation practice. The isolated-word synthesiser can be used to generate pronunciation examples for pronunciation training applications (for example, the pronunciation exercises created by the Institute of the Estonian Language[9], and the Estonian pronunciation training app SayEst[10] (Malmi et al., 2022a), which currently use human-recorded pronunciation examples), electronic and online teaching materials (e.g., the Estonian Language E-Course Keeleklikk[11]), classroom practices and self-study. For instance, unlike the other Estonian TTS systems, the isolated-word synthesiser allows for a controlled synthesis of minimal pairs differing only in palatalisation, quantity, lexical stress, or the presence/absence or location of a compound boundary, which is useful in the practice of the production and perception of these phonological features of Estonian.

Speech therapy exercises. The isolated-word synthesiser can also be used in speech therapy applications like Kõneravi.ee[12], where speech therapists can utilise existing exercises as well as create new ones. The available pronunciation and perception exercises use units at the phoneme, word, phrase, and sentence levels, words being the most frequently used perception or pronunciation units. So far, human-recorded audio examples have been used, which means that in order to create new exercises, the users must record the audio examples themselves or use examples from a limited speech database.

5 Conclusions and future work

The paper described the development, features, evaluation and use cases of the Estonian isolated-word TTS synthesiser (Kiissel, 2024 and https://elo.eki.ee/yksiksona/). The synthesiser allows the user to generate correctly pronounced isolated words and multi-word units by manually specifying the diacritics for third quantity, palatalisation, lexical stress and compound boundaries. The synthesiser generates isolated words with an appropriate utterance prosody and high phonetic quality, being thus suitable for language pedagogical and speech therapeutic applications.

Future steps include the improvement of the user-friendliness of the user interface. To help users insert the necessary diacritics, automatic tagging of the orthographic input text will be added, with multiple outputs for homographs among which the user can choose.

A second line of future work will be the development of a similar TTS application for longer texts, enabling the user to correct parsing and disambiguation errors that cause pronunciation errors.

Finally, we will employ more advanced TTS techniques to train additional isolated-word synthesisers.

[8] https://sonaveeb.ee/wordgame?uilang=en
[9] https://sonaveeb.ee/pronunciation-exercises/#/
[10] Available in Google Play store https://play.google.com/store/apps/details?id=mobi.lab.sayest&pli=1
[11] https://www.keeleklikk.ee/index_en.html
[12] https://koneravi.ee/

Acknowledgments

This study was supported by the National Programme for Estonian Language Technology 2018–2027 and by the basic governmental financing of the Institute of the Estonian Language from the Estonian Ministry of Education and Research.

References

Heiki-Jaan Kaalep and Tarmo Vaino. 2001. Complete Morphological Analysis in the Linguist's Toolbox. In Anu Nurk, Tõnu Seilenthal, and Triinu Palo, editors, *Congressus Nonus Internationalis Fenno-Ugristarum, 7.-13.8.2000 Tartu: Pars V*, Congressus Nonus Internationalis Fenno-Ugristarum, 7.-13.8.2000 Tartu, pages 9–16. Eesti Fennougristide Komitee.

Mari-Liis Kalvik and Liisi Piits. 2019. Sõna esinemissagedus ja tähenduste eristamise vajadus häälduse mõjutajana. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 10(1):71–88.

Indrek Kiissel. 2024. Merlinil põhinev üksiksõnade kõnesüntesaator [Merlin based Estonian isolated word speech synthesizer]. https://github.com/ikiissel/mrln_et_iw.

Margit Langemets, Indrek Hein, Madis Jürviste, Jelena Kallas, Olga Kiisla, Kristina Koppel, Külli Kuusk, Tiina Leemets, Sirje Mäearu, Tiina Paet, Peeter Päll, Maire Raadik, Lydia Risberg, Tuuli Rehemaa, Hanna Tammik, Mai Tiits, Katrin Tsepelina, Maria Tuulik, Udo Uibo, et al. 2023. *EKI ühendsõnastik 2023. [The EKI Combined Dictionary]*. Eesti Keele Instituut. https://sonaveeb.ee.

Anton Malmi, Katrin Leppik, and Pärtel Lippus. 2022a. SayEst - mobiilirakendus eesti keele hääldusе harjutamiseks. *Oma Keel*, 2:85–88.

Anton Malmi, Pärtel Lippus, and Einar Meister. 2022b. Articulatory properties of Estonian palatalization by Russian L1 speakers. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 13(2).

Einar Meister and Lya Meister. 2014. L2 production of Estonian quantity degrees. In *Speech Prosody 2014*, pages 929–933. ISCA.

Helle Metslang, Mati Erelt, Külli Habicht, Tiit Hennoste, Reet Kasik, Pire Teras, Annika Viht, Eva Liina Asu, Liina Lindström, Pärtel Lippus, Renate Pajusalu, Helen Plado, Andriela Rääbis, and Ann Veismann. 2023. *Eesti grammatika*. Tartu Ülikooli Kirjastus.

Meelis Mihkla, Einar Meister, and Arvo Eek. 2000. Eesti keele tekst-kõne süntees: grafeem-foneem teisendus ja prosoodia modelleerimine. In Tiit Hennoste, editor, *Arvutuslingvistikalt inimesele*, Tartu Ülikooli üldkeeleteaduse õppetooli toimetised, pages 309–320. Tartu Ülikooli kirjastus, Tartu.

Ülle Viks. 1992. *Väike vormisõnastik: Sissejuhatus ja grammatika*. Eesti Teaduste Akadeemia, Keele ja Kirjanduse Instituut.

Zhizheng Wu, Oliver Watts, and Simon King. 2016. Merlin: An Open Source Neural Network Speech Synthesis System. In *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, pages 202–207. ISCA.