Proceedings of SIGUL2022 @LREC2022, pages 10–15
Marseille, 24-25 June 2022

An Open Source Web Reader for Under-Resourced Languages

Judy Y. Fong, Þorsteinn Daði Gunnarsson, Sunneva Þorsteinsdóttir, Gunnar Thor Örnólfsson, Jon Gudnason
Reykjavík University
Menntavegur 1 - Reykjavík Iceland
judy@judyyfong.xyz, thorsteinng@ru.is, sunnevatorstein@gmail.com
gunnaro@ru.is, jg@ru.is

Abstract

We have developed an open source web reader in Iceland for under-resourced languages. The web reader was developed due to the need for a free and good quality web reader for languages which fall outside the scope of commercially available web readers. It relies on a text-to-speech (TTS) pipeline accessed via a cloud service. The web reader was developed using the Icelandic TTS voices Alfur and Dilja, but could be connected to any language which has a TTS pipeline. The design of our web reader focuses on functionality, adaptability and user friendliness. Therefore, the web reader's feature set heavily overlaps with the minimal features necessary to provide a good web reading experience while still being extensible enough to be adapted to work for other languages, high-resourced and under-resourced. The web reader works well on all the major web browsers and has a Web Content Accessibility Guidelines 2.0 Level AA: Acceptable compliance, meaning that it works well for the largest user groups, people in under-resourced languages with visual impairments and difficulty reading. The code for our web reader is available and published with an Apache 2.0 license at https://github.com/cadia-lvl/WebRICE, which includes a simple demo of the project.

Keywords: text-to-speech, web reader, accessibility, human-computer interaction

1. Introduction

The proposed open source web reader for under-resourced languages is a language technology tool for everyday users. Web readers are added to websites to let visitors listen to the content of the webpage instead of reading it. They are analogous to audiobooks made from ebooks. However, audiobooks are generally manually recorded and edited, and are labour-intensive to produce. Manually rendering text on websites into speech at scale is generally not viable. Thus, web readers use automatic text-to-speech (TTS) voices produced for the target language. In this way, web readers are a scalable way to make a website more accessible.

1.1. Language Technology Tools

Language technology (LT) tools such as web readers are scalable, which benefits under-resourced languages. It means with less effort and few resources, a wide audience can still be reached, like how web readers can allow a single TTS voice to be used on any given number of websites. One of the main goals of language technology development is to facilitate the use of natural language in today's digital age. A web reader achieves this by allowing users to listen to a given website in addition to reading it. More importantly, people visit websites in their under-resourced language every day. These smaller language communities, like Icelandic, often lack resources. In some cases, national governments can counteract this by implementing national language technology initiatives. These initiatives are often crucial to bring an under-resourced language into the digital age. For example, the Estonian language is now viable in the digital age through a government initiative described by (Meister and Vilo, 2008). With this inspiration and knowledge, the Icelandic government has implemented the five year language technology programme for Icelandic as described by (Nikulásdóttir et al., 2020), to bring the Icelandic language and the digital age together. Nikulásdóttir et al. (2021) enumerates the language technology tools and datasets created by this initiative and hosted at CLARIN-IS. One of the core areas of this initiative is text-to-speech.

1.1.1. TTS technologies

The research and development for each part of a typical TTS pipeline is considered in the Icelandic programme mentioned by Nikulásdóttir et al. (2020). For example, collecting data as in (Sigurgeirsson et al., 2020) and creating the free and open Talrómur and Talrómur 2 TTS datasets as mentioned by (Sigurgeirsson et al., 2021) and (Gunnarsson, Þ. et. al., 2021). Model training recipes have been developed for the datasets, both for unit selection[1] and neural network-based TTS methods[2][3]. TTS models from (Gunnarsson Þ. et. al., 2022) have been trained and published. Important TTS pre-processing steps like text normalization in (Sigurðardóttir et al., 2021) and grapheme to phoneme conversions as in (Nikulásdóttir, A. et. al., 2022) are also considered. A TTS web service has been developed[4], which allows anyone to host their own TTS voices.

[1] https://github.com/cadia-lvl/unit-selection-festival/

[2] https://github.com/cadia-lvl/FastSpeech2

[3] https://github.com/cadia-lvl/espnet

[4] https://github.com/tiro-is/tiro-tts

These serve as a template for providing a TTS web service in any under-resourced language. An instance of that service has been made accessible[5] as a result of the project. This allows the development of TTS applications such as the main subject of this paper: the web reader.

1.2. Other Web Readers

Under-resourced languages usually attract less commercial interest than better resourced ones. One of the reasons can be the smaller size of the language community. For example, the Icelandic language has only a few hundred thousand speakers which make up the whole market for Icelandic LT solutions. That is a tiny fraction of the millions or billions of speakers and users, which heavily resourced languages like French and English have.

This smaller commercial interest was noticeable when researching existing web readers used in the world and in Iceland. Our research delved into NaturalReader, Read Aloud, and ReadSpeaker. NaturalReader[6] seems to only offer English as it does not mention multilingual support, only a variety of voices. Also, it is a proprietary software solution. Read Aloud is offered only as a web browser extension. It is not offered as a web reader embedded into websites. It is a free in-browser web reader which connects to various proprietary TTS cloud service providers. While the extension itself is free, listening to Icelandic neural voices is only possible through paid services. Finally, ReadSpeaker[7] does support Icelandic directly but it is also a proprietary option. In addition to web readers, we also looked into screen readers such as Ivona, ClaroRead, and JAWS[8]. Screen readers are installed directly onto one's operating system and can read most text on a computer or mobile device. However, screen readers are operating system dependent and are out of scope for developing an open source web reader despite the overlap. The research results are twofold: first and foremost many web readers are commercial and second that they offer limited or no support for Icelandic, nor other under-resourced languages. The most widely adopted web reader on Icelandic websites is ReadSpeaker, built with heavy involvement from the Icelandic community a decade ago. To provide users of under-resourced languages a nearly seamless experience when using our web reader on websites, compared to proprietary readers, it would be best for our web reader to offer the same core features.

1.3. An Open Source Web Reader

As mentioned previously, under-resourced languages attract less commercial interest. In order to get international companies to implement tools for these languages, TTS language resources and tools must be open, standardized, and accessible. Only then will the needs of the under-resourced language community and commercial interests be aligned. The same is true for smaller under-resourced language companies, public entities, and individuals who usually have limited resources. So they cannot include commercial web readers. Having these language resources open and freely available also means that language technology research is more likely to be done. The result should be more websites with web readers. It is an important accessibility feature for websites to have a web reader, especially popular and required websites such as for the government, schools, and the media. Without a web reader, many under-served communities struggle to get equal access to information and events. Therefore, an open source web reader is crucial for under-resourced languages because it gets text-to-speech technology into the hands of language users immediately.

1.4. Language support

To improve an under-resourced language's chance of being included in the commercial offerings of international companies' technologies, it would appear that the most feasible option would be to make the language resources and other tools open and accessible enough for them to be incorporated easily. One way is to use the same standards of data and tools as used in these companies. This is what we have done to make sure our web reader and TTS web service support Icelandic. Our aim in making the web reader was to reach the largest possible internet audience in Iceland, meaning be good for both users (listeners) from the under-served communities and the general Icelandic population. Our open source web reader's development goal is to work with any natural language, under-resourced and highly-resourced alike, and with any TTS cloud services available. In the Icelandic case, the most popular commercial TTS cloud service offering Icelandic during initial development was Amazon Polly[9], due to it being the only TTS web service available directly for producing spoken Icelandic. But now Icelandic is also offered on two other platforms, Google[10] and Microsoft Azure[11]. Having a selection of voices is important for users and companies to choose the voice that best reflects themselves. This is why the web reader is capable of connecting to TTS cloud services from different companies.

Now that the web reader infrastructure is provided, it can be connected to TTS web services in any language. Due to the default design of our web reader, it works with Icelandic currently. But it can easily be changed to use a TTS web service from any other under-resourced language. Thus, machine learning engineers can focus

[5] https://tts.tiro.is/

[6] https://www.naturalreaders.com/index.html

[7] https://www.readspeaker.com/

[8] https://www.freedomscientific.com/products/software/jaws/

[9] https://aws.amazon.com/polly/

[10] https://cloud.google.com/text-to-speech/

[11] https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech/

solely on providing great TTS voices behind a standard HTTP POST request.

1.5. Overview

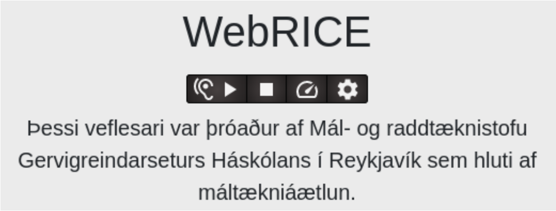
The following is a summary of our software design, implementation, and tests. It consists of requirements and features for both the web reader and integration requirements with TTS web services and websites. The software development was categorized into required and optional. In this paper, we will mainly be talking about our required features.

2. Web reader

When developing software, the first steps are to identify the core features, to understand how a web reader is used and to understand what differentiates a good web reader from a poor web reader. People rarely use bad web readers. Our web reader was developed in consultation with the target user groups, which should translate to a good user experience. More information about the target user groups is discussed later.

The web reader's visual design focus is on a high quality suite of buttons as seen in Figure 1. The web reader must be able to play, pause, resume, speed up, slow down, stop and restart audio for the given text. The buttons must also be large and visible. There are several constraints. Functionality must be intuitive; for example, the settings module must close if a user clicks away from the settings module. The web reader must work on even the longest texts. It needs to be mobile friendly, as most users browse the web on mobile devices. As the basic user interface needs to meet the "WCAG 2.0 Level AA: Acceptable compliance" accessibility standards, the keyboard interface is implemented in two ways. First is the standard keyboard interface where users or screen readers can tab through the buttons on the web reader, like on any accessible website. Second are shortcut keys on each button directly. These shortcut keys are pretty similar to the ones offered by the most popular Icelandic web reader. If a user wants to use the shortcut keys, they can read the instructions on the web reader's help menu. The web reader is also customisable, like setting reading speed. Good documentation is essential. The final requirement is it must connect to a TTS web service. To meet the need for an open source web reader, we have built a web reader whose primary focus is synthesizing text and playing audio for any language. Our web reader meets all the aforementioned requirements.

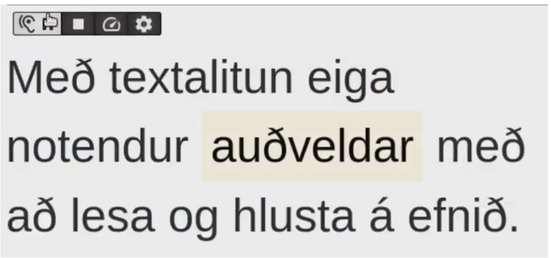
In short, our web reader consists of multiple modules. There are the button modules: play/pause, stop, speed, and settings. Other modules are highlighting, speech manager, and client store manager. Highlighting handles all the highlighting features. Text can be highlighted while users listen to the generated speech, as shown in Figure 2. Text highlighting is possible using time alignments (speech marks) to the generated speech. The speech manager module interacts with the TTS web services, and the client store manager handles all the user settings and preferences across multiple sessions.



WebRICE

Þessi veflesari var þróaður af Mál- og raddtæknistofu Gervigreindarseturs Háskólans í Reykjavík sem hluti af máltækniáætlun.

Figure 1: The play (including the ear), stop, speed, and settings buttons with some text below. The pause button replaces the play button during playback.



Með textalitun eiga notendur auðveldar með að lesa og hlusta á efnið.

Figure 2: A word being highlighted as the audio is played. The English translation of the text is as follows: With text highlighting users have an easier (highlighted word) time reading and listening to content.

Then, we have several workflows. In the preliminary workflow, the user first loads a website containing our web reader. Then, the buttons are created within the HTML tag with the web reader's ID. Next, any saved settings from a previous session are loaded from client storage. Finally, the text is extracted from the website. The main workflow starts when a user presses or selects play. The web reader fetches the generated speech and speech marks from the TTS web service. Then, we check if the user has text highlighting enabled. If so, then that is applied to the text displayed to the user.

3. Integration

Our web reader has been developed with integration heavily in mind. It has been turned into a webpack library, which can be embedded to websites with a single pre-compiled JavaScript link. Customizing the look of the web reader can be done outside of the web reader's code base, meaning developers can customize the color palette of the web player easier. A business can customize our web reader to seamlessly integrate the web reader into its own brand experience. This experience is not readily offered by any other web readers we have found on the internet.

3.1. TTS web services

Our design depends on having a separate TTS web service. The reasoning is two-fold. First, to allow our web reader to connect to multiple TTS web services. Generally, users will not use web readers with long loading times and poor quality voices. This allows users a greater choice in TTS voices, providers, and natural languages. As TTS web services are provided as standard HTTP POST requests, it would be easy to switch out the current TTS web service with TTS web services in other languages or from other providers. For example, people from under-resourced languages can connect it to their TTS cloud services and deploy the web reader for their language. Second, this separates the TTS development from the web reader development, allowing us to use both the best voices and the best user interface. But in order to connect to a variety of TTS web services, there must be a common minimum feature set that our web reader has to support: good quality voices, low latency, and speech synthesis markup language (SSML) support. Also, it should offer speech marks, which are time alignments between text and generated speech used for highlighting text.

3.1.1. TTS models

Our web reader for under-resourced languages' default integration is with a TTS web service built from Tiro's open source code repository[12]. The underlying TTS web service provides four voices in total: Karl, Dóra, Álfur and Diljá.

The data for the Álfur and Diljá voices come from Sigurgeirsson et al. (2021)'s Talrómur corpus. Two iterations of these voices have been developed. The first iteration was created from the FastSpeech 2 implementation[13] (Chien et al., 2021) as specified in (Ren et al., 2021) and has been released to the public by (Gunnarsson Þ. et. al., 2022). These models are the first publicly available open-source TTS models for Icelandic. An accompanying MelGAN[14](Kumar et al., 2019) vocoder trained on both all the voices from (Sigurgeirsson et al., 2021) for Álfur and only on Diljá for Diljá is used to synthesize the voices. Text normalization was very naive and consisted only of removing all punctuation marks and skipping all numbers. A Sequitur[15] (Bisani and Ney, 2008) grapheme-to-phoneme converter, developed by (Nikulásdóttir et al., 2018), was used for phonetic transcription[16].

We observed significant issues with these initial models, both originating in the acoustic models and the vocoder models as well as the lack of text preprocessing. Phones would often be mispronounced and noise inserted where silence would be expected. Furthermore, significant vocoder artefacts are present, especially in the Diljá voice. Thus, a second iteration of the Álfur and Diljá voices was created using the ESPNet toolkit's[17] implementation of FastSpeech 2 (Hayashi et al., 2020) and trained on the same voice data as before. Additionally, both a Parallel WaveGAN (Yamamoto et al., 2020) and a multi-band MelGAN (Yang et al., 2021) model were trained[18] on the entire Talrómur corpus (Sigurgeirsson et al., 2021). Whereas the Parallel WaveGAN model provides slightly better synthesis quality, the multi-band MelGAN model can generate samples much faster. The TTS web service which provides access to these models currently supports limited text normalization to improve the TTS output, e.g. by expanding abbreviations which should be read letter by letter rather than read as a word. The knowledge and experience gained from (Nikulásdóttir and Guðnason, 2019) shaped the text normalization created by (Sigurðardóttir et al., 2021) and which is used in the web service. Grapheme-to-phoneme conversion for out-of-vocabulary words is done using a LSTM encoder-decoder sequence-to-sequence model[19]. The web service is still in active development so expect the text normalization and other speech features to continue to improve.

3.2. Website Testing

To make sure our web reader works on various websites, we performed integration tests. During the initial development, the web reader was tested on three websites: our web reader's webpage, a local company's webpage and on our university's webpage. The web reader's color palette was also customized to match the websites' own colors. Now, in the later development stages, the web reader is being integrated into websites which did not have a web reader. These websites touch on many parts of society: including financial, government, and innovation organizations. Since the web reader is available as open source software, the organizations are able to perform this later stage of testing themselves.

4. User Tests

For web readers, some of the biggest end-users are under-served communities. Within Iceland, a large proportion of the dyslexic and visually impaired inhabitants often need to rely on spoken word for two reasons: to fully understand everything and to operate independently. However, web readers are not the best option for everyone. People who are blind or significantly visually impaired need a screen reader paired with Símarómur[20], an Android TTS engine which offers the same voices as our web reader: Álfur and Diljá. With language technology, these overlapping groups can in-

[12] https://github.com/tiro-is/tiro-tts

[13] https://github.com/cadia-lvl/FastSpeech2

[14] https://github.com/seungwonpark/melgan

[15] https://github.com/sequitur-g2p/sequitur-g2p

[16] https://github.com/atliSig/g2p

[17] https://github.com/espnet/espnet

[18] https://github.com/kan-bayashi/ParallelWaveGAN/

[19] https://github.com/grammatek/ice-g2p

[20] https://play.google.com/store/apps/details?id=com.grammatek.simaromur

dependently navigate the internet using screen or web readers, whichever best fits their situation.

In addition to the integration tests, we conducted user tests with The Iceland Dyslexia Association and the Icelandic Association of the Visually Impaired. The results revealed that these users primarily use smartphones with computers as a close second. Over 50% of respondents use web readers at least monthly. So they are recurring monthly users. Their strictest requirement is low latency voices; they have a lower tolerance than the general population for slow TTS due to TTS being a primary means of communication. Meanwhile, their most desired optional feature is selecting text and listening to it while following along with the highlighted text. However, the users did clarify that this is optional and not strictly necessary for an enjoyable web reading experience. From the results of our user tests, we are confident that users will like and use our web reader when it becomes available on websites with under-resourced languages like Icelandic.

We later extended the survey to people not in these groups and the percentage of users who use web readers monthly is significantly different in the two groups: over 50% of users in the original user tests versus under 25% of the general population. The general populace also favors browsing the internet on computers over mobile devices. So, the survey results show that web readers are disproportionately used by those with visual impairments in some way.

The results of the user tests mostly confirmed our initial research before development. However there were a few surprises. For example, that users would be satisfied by a simple interface. Users are also remarkably opposed to a nearly but not quite good TTS voice. But they are more forgiving when they know these voices will continue to improve.

5. Conclusion

The proposed open source web reader for under-resourced languages is now published and available with good documentation that describes the integration process for web developers. Connected to the web reader are two state-of-the-art Icelandic TTS voices, Álfur and Diljá. The codebase and demo for our web reader is licensed under Apache 2.0 on GitHub[21].

Now that the open source web reader is published, the aim is to integrate it to popular Icelandic websites and to make web readers more accessible to the public. This involves browser and content management system (CMS) extensions. Browser extensions allow anyone, tech savvy or not, to easily download, install and use them right away on any website. The stores for browser extensions are also an easy way for developers to easily distribute updates and bug fixes automatically for their users.

6. Acknowledgements

The authors thank Luke O'brien, Tiro ehf[22], Félag lesblindra á Íslandi (The Iceland Dyslexia Association)[23], and Blindrafélagið (Icelandic Association of the Visually Impaired)[24]. This project was funded by the Language Technology programme for Icelandic 2019-2023. The programme, which is managed and coordinated by Almannarómur[25], is funded by the Icelandic Ministry of Education, Science and Culture.

7. Bibliographical References

Bisani, M. and Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451.

Chien, C.-M., Lin, J.-H., Huang, C.-y., Hsu, P.-c., and Lee, H.-y. (2021). Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8588–8592.

Hayashi, T., Yamamoto, R., Inoue, K., Yoshimura, T., Watanabe, S., Toda, T., Takeda, K., Zhang, Y., and Tan, X. (2020). Espnet-tts: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7654–7658.

Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., de Brébisson, A., Bengio, Y., and Courville, A. C. (2019). Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32.

Meister, E. and Vilo, J. (2008). Strengthening the Estonian language technology. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Nikulásdóttir, A. B. and Guðnason, J. (2019). Bootstrapping a text normalization system for an inflected language. numbers as a test case. In *INTERSPEECH*, pages 4455–4459.

Nikulásdóttir, A. B., Guðnason, J., and Rögnvaldsson, E. (2018). An icelandic pronunciation dictionary for tts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 339–345. IEEE.

Nikulásdóttir, A., Guðnason, J., Ingason, A. K., Loftsson, H., Rögnvaldsson, E., Sigurðsson, E. F., and Steingrímsson, S. (2020). Language technology programme for Icelandic 2019-2023. In *Proceedings of the 12th Language Resources and Evaluation*

[21] https://github.com/cadia-lvl/WebRICE

[22] https://tiro.is

[23] https://www.lesblindir.is/english/

[24] https://www.blind.is/en

[25] https://almannaromur.is

Conference, pages 3414–3422, Marseille, France, May. European Language Resources Association.

Nikulásdóttir, A. B., Arnardóttir, Þ., Guðnason, J., Daði, Þ., Gunnarsson, A. K. I., Jónsson, H. P., Loftsson, H., Óladóttir, H., Sigurðsson, E. F., Sigurgeirsson, A. Þ., et al. (2021). Help yourself from the buffet: National language technology infrastructure initiative on clarin-is. In *CLARIN Annual Conference 2021*, page 124.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T. (2021). FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *Proceedings ICLR 2021 – 9th International Conference on Learning Representations*, Online, may.

Sigurðardóttir, H. S., Nikulásdóttir, A. B., and Guðnason, J. (2021). Creating data in Icelandic for text normalization. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 404–412, Reykjavik, Iceland (Online), May 31–2 June. Linköping University Electronic Press, Sweden.

Sigurgeirsson, A., Örnólfsson, G., and Guðnason, J. (2020). Manual speech synthesis data acquisition - from script design to recording speech. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 316–320, Marseille, France, May. European Language Resources association.

Sigurgeirsson, A., Gunnarsson, Þ., Örnólfsson, G., Magnúsdóttir, E., Þórhallsdóttir, R., Jónsson, S., and Guðnason, J. (2021). Talrómur: A large Icelandic TTS corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 440–444, Reykjavik, Iceland (Online), May 31–2 June. Linköping University Electronic Press, Sweden.

Yamamoto, R., Song, E., and Kim, J.-M. (2020). Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6199–6203. IEEE.

Yang, G., Yang, S., Liu, K., Fang, P., Chen, W., and Xie, L. (2021). Multi-band melgan: Faster waveform generation for high-quality text-to-speech. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 492–498. IEEE.

8. Language Resource References

Gunnarsson, Þ. et. al. (2021). *Talrómur 2 (21-12)*. CLARIN-IS.

Gunnarsson Þ. et. al. (2022). *Talrómur Utils*. CLARIN-IS.

Nikulásdóttir, A. et. al. (2022). *Icelandic Pronunciation Dictionary for Language Technology 22.01*. CLARIN-IS.