2024
THE JOURNAL ON TECHNOLOGY AND PERSONS WITH DISABILITIES

iSinkwe: An Application that Synchronises Text and Audio for Enhanced Reading

Rynhardt Kruger, Avashna Govender, Willem van der Walt, Ilana Wilken
Council for Scientific and Industrial Research
rkruger@csir.co.za, agovender1@csir.co.za, wvdwalt@csir.co.za, iwilken@csir.co.za

Abstract

We present iSinkwe, a system to produce synchronised accessible EPUB 3 books of text and audio. With iSinkwe, users are able to synchronise EPUB 3 publications with human-narrated or computer-generated speech, via an accessible web interface. Documents in other formats can also be converted to EPUB 3. Developed specifically to address reading barriers experienced by users with print disabilities such as dyslexia and visual impairment, iSinkwe is also of particular importance for regions with low literacy such as South Africa. This paper describes the motivation and context for its creation, the components that make up iSinkwe, a discussion on the relevance it has for the accessibility community, and how users can interact with the system. A usability study was performed on a previous iteration of iSinkwe, with mixed results. We report on the lessons we learned, and subsequent improvements to the system. Finally, we describe future work planned to extend its functionality.

Keywords

Accessible books, literacy, reading, synchronised highlighting.

Introduction

The development of accessible book formats like DAISY and EPUB 3 has revolutionised the way in which print-disabled readers access information, by extending the traditional audio book with sophisticated navigational capabilities (Engelen; Garrish). When text is included in these books, the reader also retains access to the symbolic form of the content, which allows interrogation with assistive technologies that output to text-to-speech (TTS) or a Braille display. However, combined text and audio EPUB 3 books are difficult to produce, requiring careful synchronisation of the source text and audio. Although TTS can be used to generate audio from text, it is inadequate when the text contains complex dialogue or words with uncommon pronunciations (Kuligowska et al. 234). In a multilingual context such as South Africa, TTS also struggles to accurately reproduce the code switching that occurs in local documents, that is, alternating between different languages in the same sentence (Setati et al. 128).

Discussion

Motivation and Context

With the advances in voice computing technologies, it is now possible to develop digital books augmented with speech. However, most of these developments have focused primarily on widely used languages such as English, German, and French. The reality is that little or no development of language technologies in most minority and under-resourced languages of the world has been done, especially those spoken in Sub-Saharan Africa. Yet these languages serve an equal purpose in the socio-economic development of communities where they are spoken. In such communities, literacy in English is typically lacking, and therefore human language technology solutions that only cater for languages such as English are not helpful. These language barriers are further aggravated by a digital/connectivity divide resulting in communities

being denied access to information in their home languages. The language and digital divides further exacerbate low literacy levels by isolating communities.

To achieve an equitable society, i.e., one where information exchange benefits all members of society, it is crucial that we develop solutions that tear down the language barriers that exist in such communities. By offering human language technologies in local languages, we can ensure that people who speak the minority languages can also enjoy literature, educational content, and other forms of information in an accessible format. South Africa, in particular, has a diverse linguistic landscape and ten of the twelve official languages can be classified as under-resourced. Therefore, developing solutions that support these languages is crucial for alleviating language barriers that exist in local communities in this country.

From an educational standpoint South Africa has many regions that struggle with literacy challenges. One contributing factor is learners having limited access to educational content in their home languages. Providing solutions that provide access to educational content in their home languages is a valuable tool that promotes literacy. In addition to language barriers, many learners with print disabilities such as dyslexia and visual impairment, as well as learners with low literacy also struggle with reading barriers that most human language technology solutions don’t cater for.

Proposed Solution

As a solution to the above-mentioned challenges, we present iSinkwe, a system that augments EPUB 3 publications by automated synchronisation of text with audio. By using iSinkwe, existing human-recorded audio can be synchronised with the text, to word, sentence and paragraph level. This facilitates the creation of multimodal accessible books which provides the benefits of symbolic text combined with the rich expressiveness of audiobooks produced by

human voice artists (Knox 127). iSinkwe can also utilise TTS to synthesise audio, for parts of the book for which no pre-recorded audio is available. The TTS functionality is capable of auto-language switching, based on the language tags in the HTML.

The audio from either natural sources or TTS is embedded inside the publication using EPUB 3 media overlays. This embedding is particularly suited for the South African context, where data connectivity is sporadic, and users may not have access to high quality voices provided as cloud services. However, since both the text and audio are contained within the EPUB 3, users retain the capability to explore the text of the book with their chosen screen reader (King 265). iSinkwe is also capable of generating older DAISY 2.02 publications using the DAISY Pipeline 2, to be played on older devices that users may still retain.

Interaction

iSinkwe consists of three components. iSinkwe Convert is a converter, to allow conversion of alternative document formats into EPUB 3. iSinkwe synchronize synchronizes the text of EPUB 3 documents with audio. Finally, iSinkwe Read is an EPUB 3 reader which supports media overlays.

The web user interfaces for iSinkwe Convert and iSinkwe Synchronise is presented in Figure 1 and Figure 2. In the iSinkwe Convert web user interface, the user needs to upload their document, select an EPUB 3 layout and enter the title and author of their desired EPUB 3 document. In the iSinkwe Synchronise web user interface, the user is expected to upload their EPUB 3 document, select the language of the EPUB 3 book, select the type of audio they desire (human-narrated or computer-generated) and select the voice characteristics which includes the gender and specific speaker of the voice. In the case where users selected human-narrated audio

as the audio type, they are also expected to upload a corresponding audio file for each section of the book.

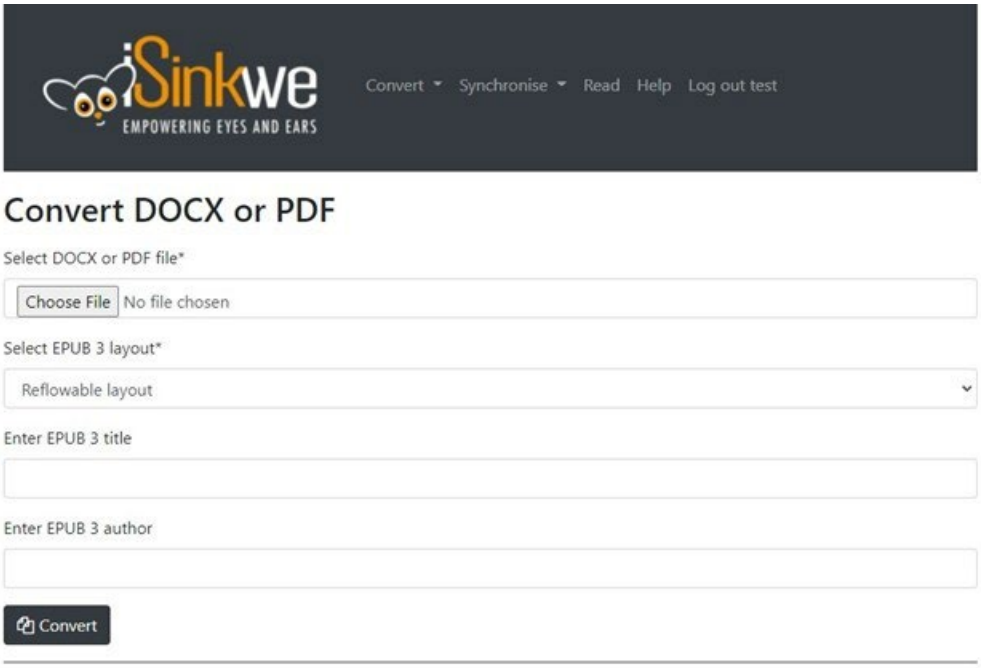

Fig. 1. Screenshot of iSinkwe Convert.

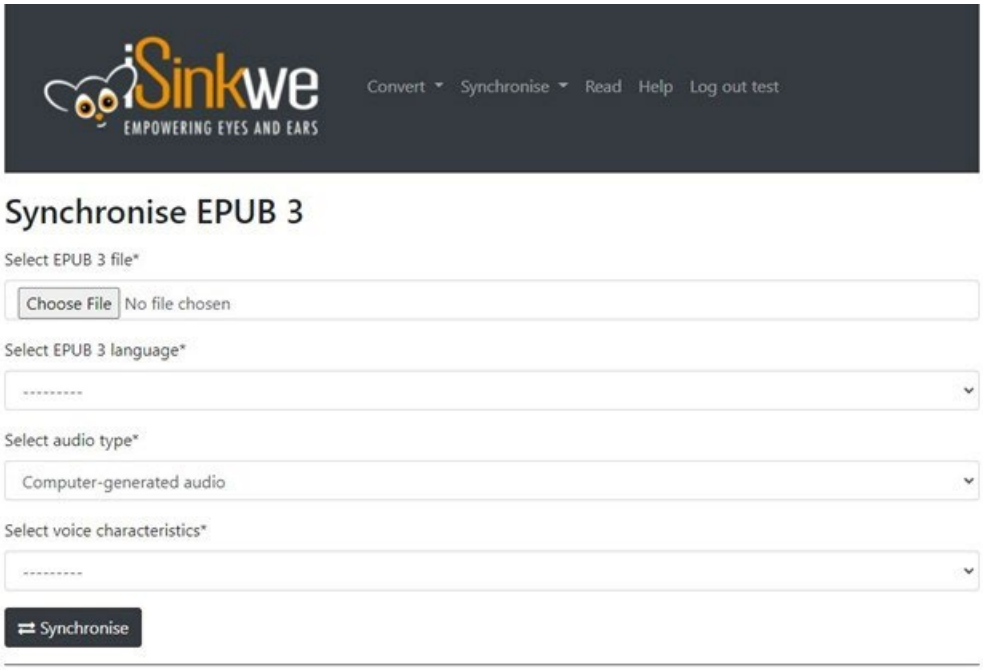

Fig. 2. Screenshot of iSinkwe Synchronise.

The mobile application user interface for iSinkwe Read is illustrated in Figure 3. In this function, the synchronised EPUB 3 created in iSinkwe Synchronise is given as input to iSinkwe Read, which can be downloaded by the user to a mobile device. In the figure, a screenshot of iSinkwe Read is presented which demonstrates the highlighted sentence feature. When a sentence is highlighted, the sentence is simultaneously read out loud using either human-narrated speech or computer-generated speech. The settings allow the user to choose whether they prefer highlighting on word, sentence or paragraph level (highlighted in black, grey and green in Figure 3 respectively).

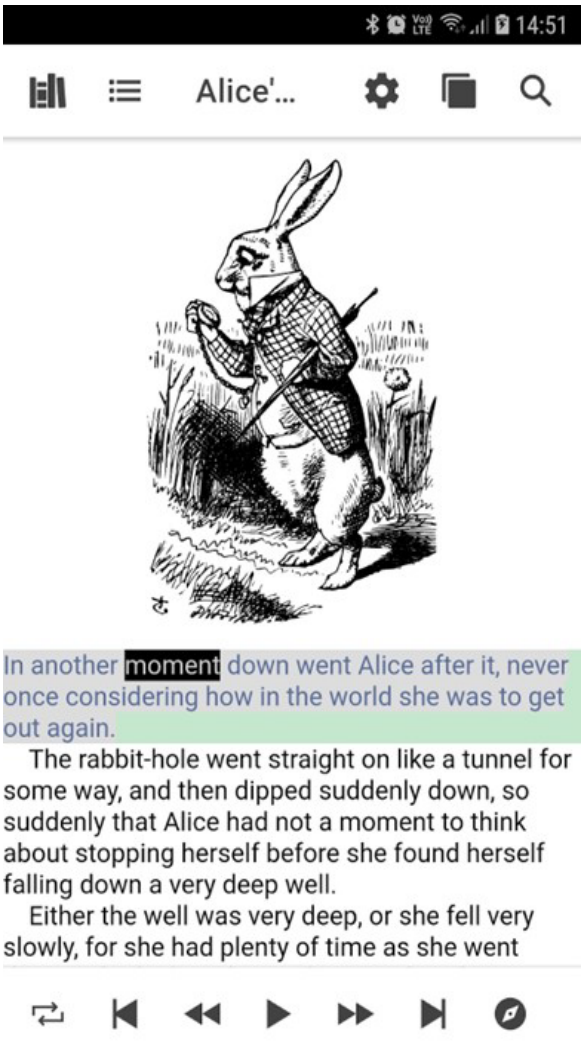

Fig. 3. Screenshot of iSinkwe Read.

Technical Description

iSinkwe does the automatic alignment by first extracting the XHTML files from the input EPUB file, then adding three levels of span tags with ID attributes to each file (paragraph, sentence and word level). If a pre-existing audio file exists for a given XHTML file or sub-section of such a file, alignment is done immediately. If no pre-existing audio is available, the text is first synthesized using text-to-speech to create an audio file. Using a dynamic time warping (DTW) aligner (Müller 69), audio offsets for the text are calculated. The output of the aligner is a SMIL file which is then added to the EPUB document according to the media overlay specification in the EPUB 3 standard.

DTW requires us to synthesize the text to an audio file and compare that with the audio file provided as input. Put another way, audio is synthesized even if pre-existing human audio is available, however, the user never hears this audio as it is only used internally to do the alignment. DTW is language dependant. A TTS engine that supports the language or languages in the book is therefore required. For South African languages, iSinkwe utilises a TTS engine also developed by our research group (Louw et al.), commercialised under the name Qfrency (CSIR). However, alignment in any other language can be supported for which a suitable TTS engine is available.

Evaluation

A previous version of iSinkwe was evaluated by a small group of diverse users. In total, ten users participated in the evaluation, comprising five blind or visually impaired users, three sighted educators for disabled learners, and two sighted members from the publishing industry. Because Covid-19 was still a realistic threat in South Africa at the time this work was performed,

we decided to conduct the evaluation virtually. However, users were also offered the option of an in-person evaluation by a team member.

To evaluate iSinkwe, users were asked to perform a number of tasks with pre-selected documents as well as with their own documents. A questionnaire with qualitative and quantitative questions was administered after the trials to gather feedback on the participants' experience with each component of the system (iSinkwe Convert, iSinkwe Synchronise, and iSinkwe Read). Questions included what users liked the most about the component, what users liked the least, whether they experienced any issues, and what they would change. Users were also asked to rate how likely they were to recommend the component on a scale from one to ten, where ten is the most likely.

Figure 4 depicts the possibility that users would recommend iSinkwe Convert to someone they know. Some of the reasons that users gave for recommending iSinkwe Convert include "it's a very functional product that can save blind people a lot of trouble once they have a text document they want to read", and "it is so user-friendly and accessible, even people with limited computer literacy would be able to use it". One user who gave a score of one stated that they do not currently see a need for EPUB 3 documents. Another stated that they also require the conversion of charts and other graphical content. Our current improvements to the system focus on the conversion of graphical mathematical content, and we hope to extend this functionality to charts and diagrams in the future.

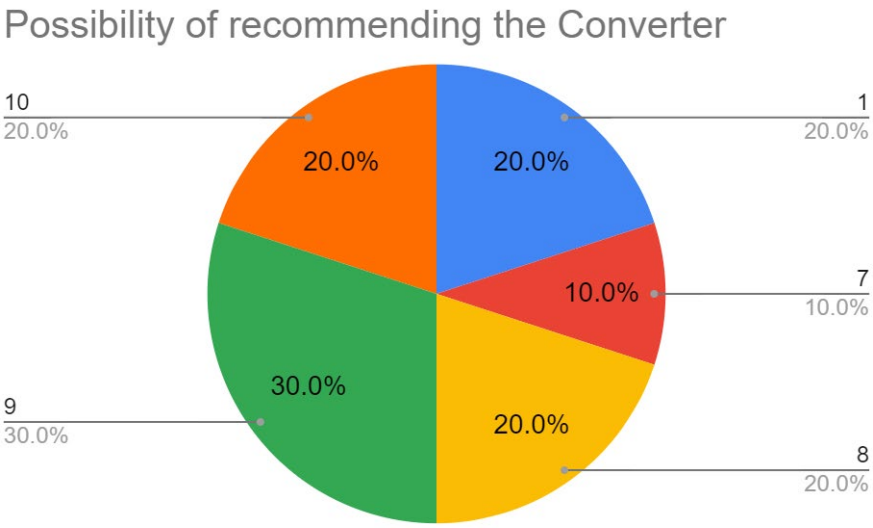

Fig. 4. The possibility of users recommending iSinkwe Convert to their friends, colleagues or family members.

Figure 5 depicts the possibility that users would recommend iSinkwe Synchronise to someone they know. Some of the aspects that users appreciated about iSinkwe Synchronise include “The clearly marked buttons and combo boxes.”, and “Having the freedom to choose the voices and type of augmentation.”. Some of the reasons why users scored iSinkwe Synchronise highly include “Although there are some teething troubles, it is very accessible and a brilliant tool for text-to-speech on documents which saves time.”, and “It’s a really functional product.”. The users who gave a score of one gave as reasons that the synchronisation takes too long, and that they found the interface rather complex with many steps to follow. Unfortunately, the speed of synchronisation is dependent on the DTW algorithm, although users are able to start the process and log off from the interface while the synchronisation takes place. Users are notified via email when the synchronisation is complete.

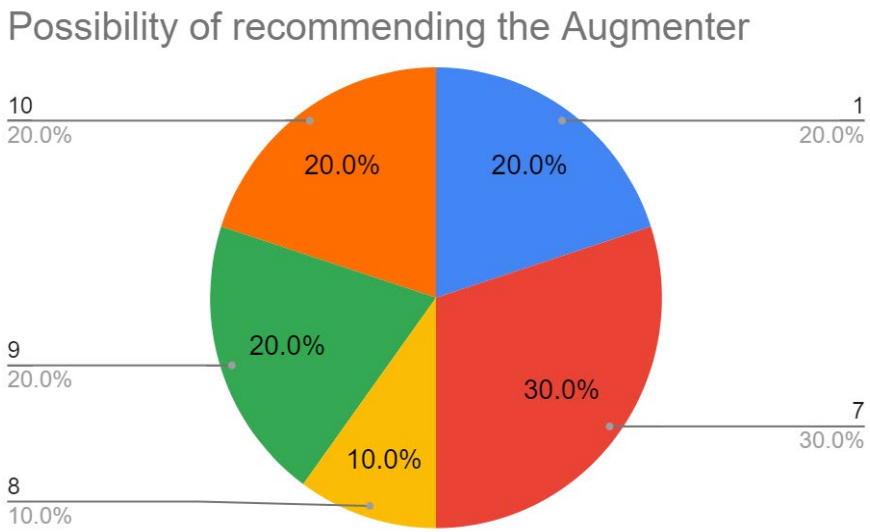

Fig. 5. The possibility of users recommending iSinkwe Synchronise to their friends, colleagues or family members.

Figure 6 depicts the possibility of users recommending iSinkwe Read to someone they know. Two users indicated that they will not likely recommend the Reader to other users (they gave a score of 1), since it "does not have any of the capabilities of the screen reader that we are already using at our school". One user gave a rating of 3, stating the Reader still needs work. The users who gave a score of 9 said the speech is clear and the controls are labelled clearly, but the navigation is confusing to new users, and also that the Reader is "very user friendly and efficient (once you find and get your book imported)".

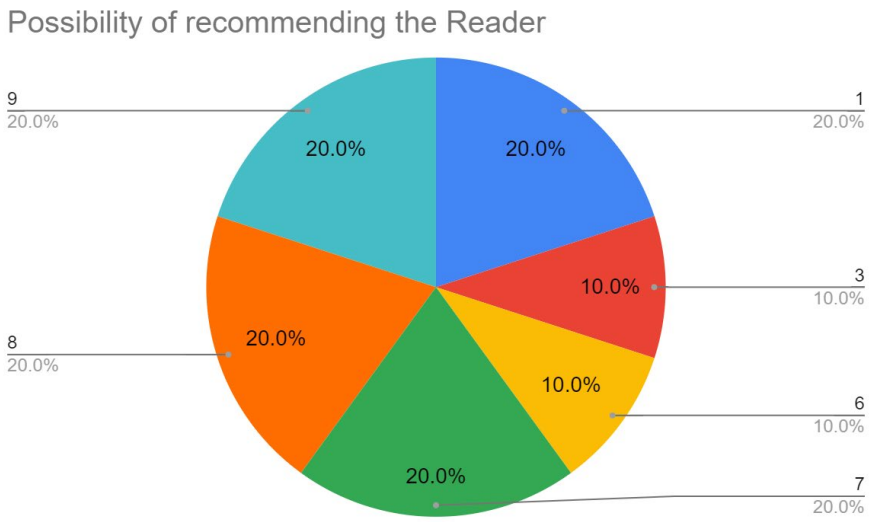

Fig. 6. The possibility of users recommending iSinkwe Read to their friends, colleagues or family members.

Overall, the blind and visually impaired users had a mostly positive experience with iSinkwe Convert and iSinkwe Synchronise, although they suggested some improvements from the web content accessibility guidelines (WCAG), which were subsequently integrated into the system. However, their experience with iSinkwe Read was mixed, with many stating that they already have ebook readers that they prefer. We therefore continue to focus on standard conforming EPUB 3 books that can be read with any reader that supports media overlays.

The members of the publishing industry were specifically interested in iSinkwe Convert and iSinkwe Synchronise, and therefore mainly used iSinkwe Read to verify the resulting documents. They had a mostly positive experience with iSinkwe and commented that iSinkwe has the potential for opening a world of reading to those who cannot or who struggle to read. In contrast, the educators had the least positive experience of the three user groups. At the time, iSinkwe did not fit their use case, since they required software that allows learners to read and write on the same document. Their experience can also be attributed to the high workload of

educators in South Africa, especially educators for disabled learners, resulting in their limited time to explore new technologies. We have subsequently created an instruction video, to assist users in familiarizing themselves with the system.

The participants also had to complete a Systems Usability Scale (SUS) questionnaire which assisted with determining the usability of iSinkwe (Brooke 189). iSinkwe obtained a SUS score of 62, which measures as a D-grade on the scale. This means iSinkwe was rated as between OK/Fair and Good. A score of 68 is deemed average. However, significant development has since taken place and a new user evaluation is needed to calculate an updated score for iSinkwe.

Future Work

Current development focuses on an iSinkwe component for automatically recognising inaccessible mathematical expressions from bitmap images using optical character recognition capable of interpreting mathematical information. By utilizing the MathJax rendering library (Cervone et al.), these expressions are augmented with MathML which the user can explore, as well as a textual description to be synchronised using the existing audio synchronisation functionality. In future work, we aim to extend this image recognition functionality to also recognise other technical information types like diagrams that occur in textbooks (Emerson and Anderson 20) and develop a method by which users may explore these objects.

Conclusion

This paper described iSinkwe, a solution for augmenting books and other electronic documents with human and/or computer-generated speech. iSinkwe accomplishes this by utilising the EPUB 3 document format with media overlays.

iSinkwe was evaluated by users, teachers and publishers, and the feedback received during the evaluation was incorporated into subsequent versions of iSinkwe as far as it was

possible and feasible. Improvements are also made when necessary to ensure a good user experience for those who use iSinkwe.

iSinkwe's core functionality is to create multimodal digital books and therefore plays a key role in accessibility to individuals with visual impairments and other print disabilities. It provides an encouraging solution for those who have reading barriers and in addition supports users who prefer to read in their home languages that are resource scarce. By synchronising text with audio, it adds an additional mode of reading to existing documents, thereby offering a multi-modal solution that allows users to read in a more interactive and engaging way which has many advantages as discussed in (Rumsey 191). Therefore, especially in the low-literacy context, iSinkwe contributes towards turning non-readers into readers, makes content available to a wider range of users and promotes a more interactive and engaging approach to literacy development.

Works Cited

Brooke, John. "SUS: a 'quick and dirty' usability scale." Usability evaluation in industry (1996): 189-194.

Cervone, Davide, Peter Krautzberger, and Volker Sorge. "Towards universal rendering in MathJax." *Proceedings of the 13th International Web for All Conference*. 2016.

CSIR. "Qfrency". *Council for Scientific and Industrial Research*. www.qfrency.com. Accessed 4 Oct. 2023.

Emerson, Robert Wall, and Anderson, Dawn. "What mathematical images are in a typical mathematics textbook? Implications for students with visual impairments." *Journal of Visual Impairment & Blindness* 112.1 (2018): 20-32.

Engelen, Jan. "E-books and audiobooks: what about their accessibility?" *International Conference on Computers for Handicapped Persons*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

Garrish, Matt. "What is EPUB 3?" *O'Reilly Media, Inc*., 2011.

King, Alasdair. "Screenreaders, magnifiers, and other ways of using computers." *Assistive Technology for Blindness and Low Vision* (2018): 265-288.

Knox, Sara. "Hearing hardy, talking Tolstoy: The audiobook narrator's voice and reader experience." *Audiobooks, literature, and sound studies*. Routledge, 2011. 127-142.

Kuligowska, Karolina, Paweł Kisielewicz, and Aleksandra Włodarz. "Speech synthesis systems: Disadvantages and limitations." *International Journal of Engineering & Technology* (UAE) 7 (2018): 234-239.

Louw, Johannes A., et al. "The speect text-to-speech entry for the Blizzard Challenge 2016." *Blizzard Challenge 2016*. Cupertino, United States of America. 16 September 2016.

Müller, Meinard. "Dynamic time warping." *Information retrieval for music and motion* (2007): 69-84.

Rumsey, Francis. "Audio in multimodal applications." *Journal of the Audio Engineering Society* 58.3 (2010): 191-195.

Setati, Mamokgethi, et al. "Incomplete journeys: Code-switching and other language practices in mathematics, science and English language classrooms in South Africa." *Language and education* 16.2 (2002): l 128-149.

"Web content accessibility guidelines (WCAG) 2.1". *World Wide Web Consortium*. www.w3.org/TR/WCAG21/. Accessed 5 Dec. 2023.