Applications in Accessibility of Text-to-Speech Synthesis for South African Languages: Initial System Integration and User Engagement

Georg I. Schlünz
Human Language Technology
Research Group
CSIR Meraka Institute
Pretoria, South Africa
gschlunz@csir.co.za

Ilana Wilken
Human Language Technology
Research Group
CSIR Meraka Institute
Pretoria, South Africa
iwilken@csir.co.za

Carmen Moors
Human Language Technology
Research Group
CSIR Meraka Institute
Pretoria, South Africa
cmoors@csir.co.za

Tebogo Gumede
Human Language Technology
Research Group
CSIR Meraka Institute
Pretoria, South Africa
tgumede@csir.co.za

Willem van der Walt
Human Language Technology
Research Group
CSIR Meraka Institute
Pretoria, South Africa
wvdwalt@csir.co.za

Karen Calteaux
Human Language Technology
Research Group
CSIR Meraka Institute
Pretoria, South Africa
kcalteaux@csir.co.za

Kerstin Tönsing
Centre for Augmentative and
Alternative Communication
University of Pretoria
Pretoria, South Africa
kerstin.tonsing@up.ac.za

Karin van Niekerk
Centre for Augmentative and
Alternative Communication
University of Pretoria
Pretoria, South Africa
karin.vanniekerk@up.ac.za

ABSTRACT

Persons with certain disabilities face barriers to information access and interpersonal communication. Assistive technologies provide workaround solutions to these problems. Augmentative and alternative communication systems aid the person with little or no functional speech to speak out loud. Screen readers and accessible e-books allow a print-disabled (visually-impaired, partially-sighted or dyslexic) individual to read text material by listening to audio versions. Text-to-speech synthesis converts electronic text into artificial speech and is used as the vocalisation component in the assistive technologies. For these three use cases, we report on an initial round of system integration and user engagement of the Qfrency text-to-speech voices that provide access to synthetic speech in the South African languages.

CCS CONCEPTS

• **Human-centered computing → Sound-based input / output**; **Empirical studies in accessibility**; **Accessibility technologies**;

KEYWORDS

Accessibility, assistive technology, text-to-speech, South African languages, system integration, user engagement

ACM Reference format:
Georg I. Schlünz, Ilana Wilken, Carmen Moors, Tebogo Gumede, Willem van der Walt, Karen Calteaux, Kerstin Tönsing, and Karin van Niekerk. 2017. Applications in Accessibility of Text-to-Speech Synthesis for South African Languages: Initial System Integration and User Engagement. In *Proceedings of SAICSIT '17, Thaba Nchu, South Africa, September 26–28, 2017,* 10 pages.
https://doi.org/10.1145/3129416.3129445

1 INTRODUCTION

Interpersonal communication is a fundamental part of our daily human lives. We need to communicate with one another when we want to speak our hearts and minds, not only to be heard but also to be understood. We use different channels to convey our messages, including verbal speech, written text, and drawn images. We alternate between the roles of the speaker (writer, artist) and listener (reader, interpreter). During our turn-taking, we reflect on what each party has said (written, drawn) and we adjust our interaction to accommodate the other party in order to reach a mutual understanding.

SAICSIT '17, September 26–28, 2017, Thaba Nchu, South Africa

ACM ISBN 978-1-4503-5250-5/17/09…$15.00
https://doi.org/10.1145/3129416.3129445

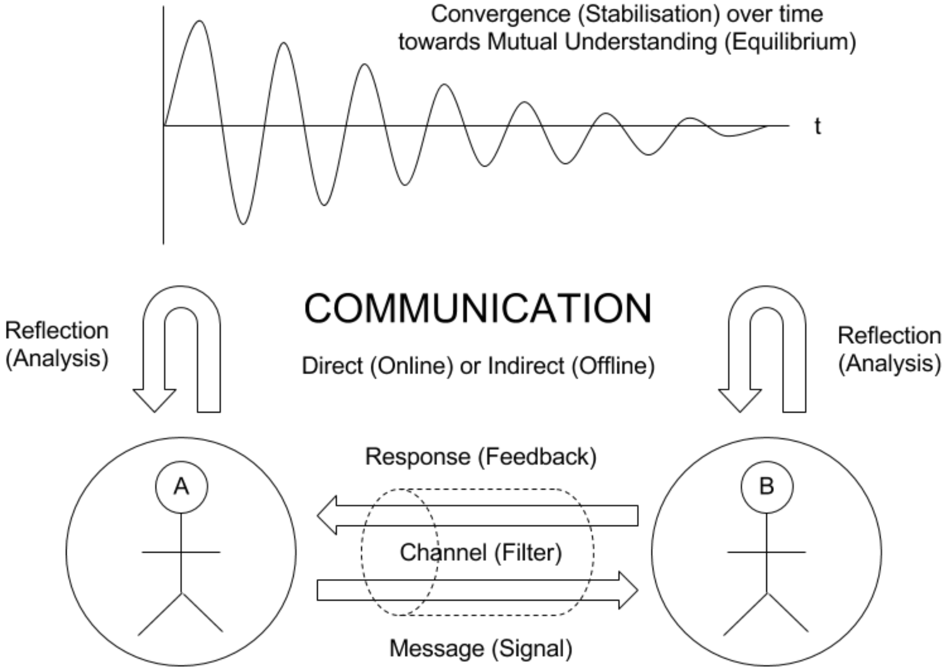

Convergence (Stabilisation) over time
towards Mutual Understanding (Equilibrium)
t
COMMUNICATION
Direct (Online) or Indirect (Offline)
Reflection
(Analysis)
Reflection
(Analysis)
Response (Feedback)
A
B
Channel (Filter)
Message (Signal)

Figure 1: Communication processes

Certain people, however, experience gross barriers to communication because of physical and/or intellectual impairments. A person whose speech is severely impaired cannot speak to convey his or her message verbally. The print-disabled (visually-impaired, partially-sighted or dyslexic) individual cannot read a book or browse the Internet in a conventional way in order to access information. "Accessibility" is an umbrella term denoting the advocacy, policies, guidelines, and solutions that seek to bridge the gaps that exist in access to infrastructure and services for persons with disabilities [11, 12]. The focus of this paper is on information and communication-related access.

Information and communication technology (ICT) is a broad enabler of this type of accessibility. Fig. 1 illustrates the link between human communication and ICT, in other words, it relates the sociological definition of communication to its technological implementation. The communication processes are drawn using an engineering analogy. The analogy borrows its diagram components and terminology (in brackets) loosely from the control system and signal processing theory [16].

Human language technology (HLT) [6], in particular, text-to-speech (TTS) synthesis [16], is a specialised subset of ICT that can play an important role in the accessibility domain. TTS can synthesise speech on behalf of the person with a natural speech impediment, as well as verbalise written text in books and on the Internet into audio to which the print-disabled individual can listen. The Human Language Technology Research Group (HLTRG) at the CSIR Meraka Institute has developed TTS voices in the South African languages and is commercialising these under the brand name "Qfrency TTS" [8].

This paper discusses the application of Qfrency TTS in three accessibility-related use cases, with the aim to showcase the fit of the technology for overcoming the aforementioned communication barriers. The next sections relate the work done for each use case, by way of sketching the background, describing the system integration and reporting on the user engagement. The paper concludes with a summary of the findings, recommendations, and future work.

2 USE CASE 1: AUGMENTATIVE AND ALTERNATIVE COMMUNICATION

2.1 Background

A severe speech impairment may be caused by motor neuron disease, cerebral palsy, a stroke, autism and many other conditions. Augmentative and alternative communication (AAC) is a field of work that addresses this communication obstacle by supplementing or replacing the person's traditional means to written and/or spoken language with assistive technology [10]. Its ultimate aims are to empower the individual with improved social skills and to equip him or her to become a respected, contributing member of society.

In partnership with the Centre for Augmentative and Alternative Communication (CAAC) at the University of Pretoria, we set out to engage with South African multilingual persons using AAC, as well as their families and service providers. The aim was to perform a baseline integration of the existing Qfrency TTS voices into a selected AAC system and evaluate the user experience. We realise that multilingual AAC system design is a complex task; therefore, concurrently, we also elicited from the end-users and service providers the desired features of AAC systems that can give access to expression in multiple languages. We are going to report on this user requirements study in detail in the near future [18, 19].

2.2 System Integration

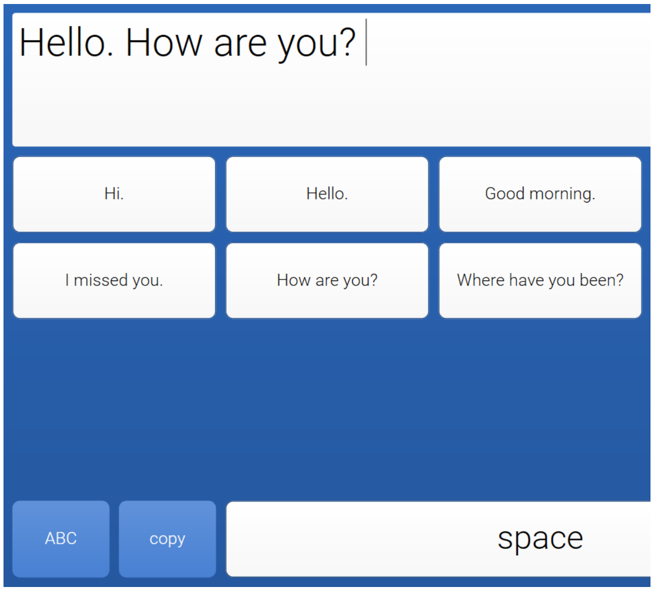
An AAC system can assist its user to produce a message using different symbols as input. The literate person may opt for text-based production, whereas the low-literate individual or child still acquiring literacy skills may prefer graphic symbol-based production. The input choices are typically categorised in cells on a grid-like layout. The level of granularity for text input can be set from single letters to individual words or to complete phrases and sentences. The graphic symbols typically represent a word or larger concept. Different grids can be set up for different combinations of symbols, with the aim to optimise sentence construction. The input can be selected via keyboard, mouse or touch screen if the end-user has the motor skills to do so, otherwise, assistive devices can be employed. Switches enable the end-user to navigate the cells in a grid by row and column, whereas eye trackers allow the end-user to focus on a cell to select it. The AAC system can be set to speak the whole message out loud with a TTS voice or the constituent parts as they are constructed.

Grid 3 from Smartbox Assistive Technology [17] is an AAC system that is sold in South Africa and used (amongst other hardware and software) for evaluation purposes by the CAAC. We successfully integrated the Qfrency TTS voices into Grid 3 that supports the Microsoft Speech Application Programming Interface (SAPI), a native speech processing interface for Windows. There was no specific need to perform a direct integration into Grid 3. We also used the customisation tool of the system to build text interfaces with sample sentences for the South African languages. Figs. 2 and 3 show screenshots of one of the standard Grid 3 text interfaces for English compared to a customised interface for isiZulu, respectively.

2.3 User Engagement

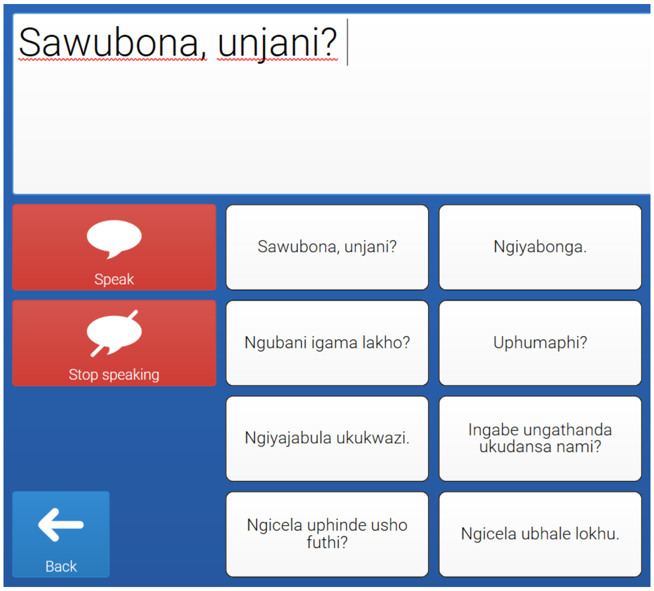
We recruited selected literate AAC end-users from a network maintained by the CAAC to do acceptance testing of the TTS voices via a facilitated questionnaire that addresses intelligibility and naturalness factors. The literacy requirement had to be enforced for the text interfaces; graphic symbol-based interfaces will be developed in future work to include non-literate end-users as well.

The intelligibility of a TTS voice is mainly concerned with the correct pronunciation of words so that the listener can understand what is being said. A respondent is typically asked to write down or type what he or she heard. The naturalness of a TTS voice is primarily determined by prosody. Prosody includes phrase breaks, sentence-level stress, and intonation, as well as word-level stress or tone. These factors can be aggregated into more general, qualitative questions to the respondent that centre around how "human-like" the TTS voice sounds. These speech qualities are important since the TTS voice "mediates" not only the meaning of the message but also the personality of the AAC user to an extent. These are important factors that contribute to whether or not the conversation between the AAC user and his or her partner is acceptable and, therefore, successful.



Hello. How are you?

Figure 2: Standard Grid 3 text interface for English



Sawubona, unjani?

Figure 3: Customised Grid 3 text interface for isiZulu

We had to simplify the traditional approaches to these scientific measures to cater for the fact that verbal and written communication is difficult for the end-users due to their speech and motor impairments. We employed multiple choices as a closed-form answering mechanism that allowed us to guide the end-users more easily through the questionnaire.

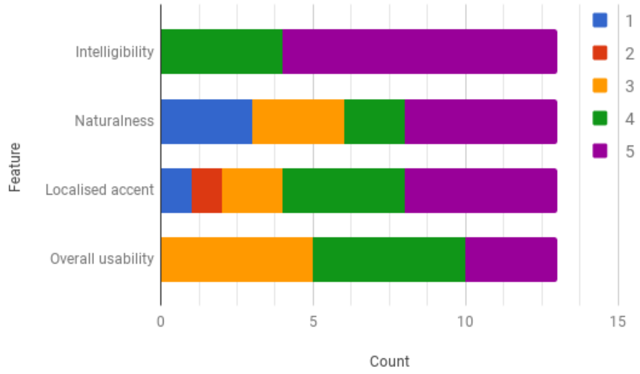


Figure 4: Aggregated TTS voice quality results

Another complication arose for the intelligibility tests. The reason why written transcription is traditionally used instead of a verbal answer to indicate understanding is to prevent the scenario where the listener could simply be parroting back what he or she is hearing. Although the end-users involved in this use case understand their non-English native languages, very few can actually spell in them. Among the reasons are their primarily English schooling and their use of the historically English-only AAC systems. We, therefore, leveraged the fact that every participant in this evaluation is literate in English and redesigned the intelligibility transcription as a multiple choice translation task between the native language and English. We used 10 native/non-English sentences and note that the sentences were not shown to the end-user, but synthesised with the TTS voice and played back.

The responses to the questionnaire are listed in Table 2 and illustrated in Fig. 4. The table headings, figure labels and values refer to the questions and answer options in Table 1.

Intelligibility scored high, which is an important feat since understandable synthetic speech is a foundational requirement of a usable TTS voice. The naturalness ratings are more spread out between the two poles of robotic and human-like synthetic speech. It remains an ongoing research topic. The current ability of the TTS voices to exhibit localised, mother-tongue accents is deemed acceptable to good. It is correlated with naturalness and should improve along with the latter. Overall, the respondents deem the voices acceptable to use. In particular, those who could not fit the voices to their personalities nonetheless chose the pragmatic option of using the voices until more personalised versions become available (six opted “Yes”, seven “Until” and zero “No”).

3 USE CASE 2: SCREEN READING

3.1 Background

At present, electronic documents that need to be viewed on a computer as part of our daily tasks mainly employ the modality of text to represent a message. Print-disabled and low-literate individuals are often not afforded access to such content. Furthermore, the global progression of the World Wide Web has enabled us to transact with (publish and retrieve) information on a larger scale than ever before. However, content creators on the Web do not always make provision for more convenient access to their information by persons who are print-disabled and/or low-literate. This widens the gap between these communities and the mainstream world even more. A screen reader is an assistive technology that, amongst other functionalities, enables navigation through the structure of electronic documents and web pages by means of a keyboard, as well as the reading of the content of such documents and web pages in the alternative modality of audio, using TTS.

This use case entailed the integration of the Qfrency TTS voices into an open-source screen reader. In preparation, we investigated the current web accessibility standards of the World Wide Web Consortium (W3C), namely the Web Content Accessibility Guidelines (WCAG) [2] and the Web Accessibility Initiative Accessible Rich Internet Applications (WAI-ARIA) [4]. These documents contain standards and best practices that describe how a web page should be developed to provide an accessible interface to end-users who use assistive technologies such as screen readers. We collaborated with two institutions to test the user experience, namely a network of public ICT centres that provide ICT-related services to end-users who are print-disabled and/or low-literate, and a private learning institution that offers formal qualifications to blind students in the fields of office administration, management, and marketing. The local languages of South African English, Afrikaans and isiXhosa are employed at these venues.

3.2 System Integration

The ICT centres and private learning institution use Windows computers. The former furthermore required that the screen reading solution must be executable in a locked-down, thin-client environment, that is without installation or administrative rights on light-weight target computers. This required packaging the solution in a portable, autonomous format, with no external dependencies besides that found on a typical Windows computer.

We selected the open-source NVDA screen reader that is not only free but also allows direct integration of the Qfrency TTS voices using a custom driver that we developed. The NVDA screen reader supports SAPI too, though the locked-down requirement necessitated the custom driver bundled with NVDA. The driver serves two functions: it acts as a bridge between the screen reader and the speech synthesiser, and it maintains the configuration of the synthesiser. By acting as a bridge, it allows the NVDA screen reader to produce synthesised speech in the local languages. It also maintains the configuration of the speech synthesiser, including the current language and speaking voice. It can change the language when the screen reader encounters text written in another language for which the synthesiser has a voice available. This allows the driver to switch languages automatically when, for instance, isiXhosa text is encountered on an English web page as long as the specific text is properly tagged in the HTML markup. The interaction between the NVDA screen reader and Qfrency TTS can be described in the following steps and visualised by the flow diagram in Fig. 5:

(1) The NVDA screen reader receives textual information from the user's computer. This textual information is usually either the text currently displayed on the screen from a document or web page, or the text typed on the keyboard.
(2) The NVDA screen reader passes this text on to the driver.

Table 1: Legend for AAC evaluation results

Questions	Responses
"Intelligibility": How many words spoken by the voice did you understand?	1 = I understood none of the words 2 = I understood some of the words 3 = I understood half of the words 4 = I understood most of the words 5 = I understood all of the words
"Naturalness": How much does the voice sound like a human?	1 = The voice sounds completely robotic 2 = The voice sounds a little robotic 3 = The voice sounds acceptable 4 = The voice sounds a little human 5 = The voice sounds completely human
"Localised accent": How much does the voice sound like a mother tongue speaker?	1 = The voice has a heavy foreign accent 2 = The voice has a slight foreign accent 3 = The voice sounds acceptable 4 = The voice has a slight mother tongue accent 5 = The voice has a proper mother tongue accent
"Personality fit": Does the voice fit your personality?	Yes = Yes, this voice is perfect! No = No, I want another voice that suits my personality better Until = I will use this voice until a more personalised voice becomes available
"Overall usability": What is your overall impression of the voice in terms of usability?	1 = The voice is very bad, I can't use it at all 2 = The voice is bad, I don't want to use it 3 = The voice is satisfactory, I will use it out of necessity 4 = The voice is good, I want to use it 5 = The voice is excellent, I definitely want to use it

Table 2: Detailed AAC evaluation results

Language	Gender	Sentence Score	Intelligibility	Naturalness	Localised Accent	Personality Fit	Overall Usability
Afrikaans	Female	10	5	5	5	Yes	5
isiXhosa	Female	10	5	4	5	Until	3
isiXhosa	Female	10	5	1	3	Until	3
isiXhosa	Male	10	4	1	3	Yes	4
isiZulu	Female	10	5	4	4	Until	3
isiZulu	Male	10	4	3	4	Until	4
isiZulu	Male	10	5	3	2	Until	4
isiZulu	Male	10	4	5	4	Until	4
Setswana	Female	10	5	5	5	Yes	5
Setswana	Female	10	5	5	5	Yes	5
Setswana	Female	10	5	3	4	Until	3
SiSwati	Female	10	5	5	5	Yes	4
Tshivenda	Male	10	4	1	1	Yes	3

(3) The driver sends the text to the speech synthesiser.
(4) The synthesiser sends auditory information back to the driver.
(5) The driver then outputs this auditory information through the computer's speaker system.

3.3 User Engagement

We used a questionnaire to collect information on the demographics of the respondents and their current use of assistive technologies, including their preferences with regard to TTS voices. We also administered a questionnaire to the managers of the ICT centres, to obtain information on the assistive technologies provided at the centres. We then facilitated a focus group discussion to obtain

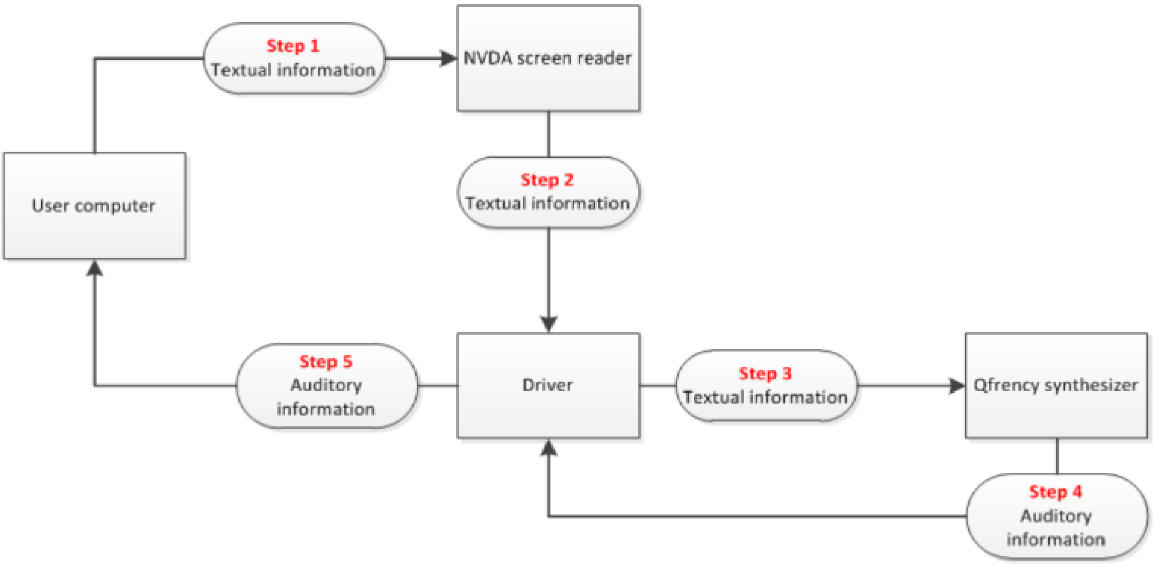


Figure 5: NVDA screen reader system flow diagram

qualitative information from the respondents on their user experience with the screen reading solution as a whole, including the ease-of-use, and whether or not the quality of the Qfrency TTS voices was acceptable for the task at hand.

We installed the solution at the public ICT centres and afforded the end-users the opportunity to use the solution over an extended period of time. Being aware that technology uptake is generally slower in developing regions [1, 7], we engaged extensively with the provincial ICT managers, centre managers, and end-users, through presentations, discussions, and training on the use of the technology. We also marketed the availability of the solution by means of flyers, social media, and word-of-mouth. While this certainly resulted in raising awareness on assistive technologies, we found that the technology literacy levels of end-users severely impeded the uptake of the solution. In addition to addressing slow responses to new technology, increasing access to information through the use of ICTs must therefore also consider the readiness levels of the end-user community into which the technology is being introduced.

Eleven end-users from three public ICT centres participated in the evaluations. We observed that, in order to use the screen reading solution effectively, the end-users would need basic knowledge on how to use a computer and keyboard. Furthermore, as no assistive technology had previously been available at these ICT centres, the end-users were apprehensive in trying out the solution. As a result, we could not conduct a detailed evaluation and it had to suffice that we introduce the solution to the respondents and allow them the opportunity to learn how to use it. Informal feedback suggested that the respondents found the idea of a screen reader in the local languages fascinating, as it gave them access to information that they had not had before.

At the private training institution, seven blind end-users were engaged in the evaluations. In contrast to the situation at the public ICT centres, these respondents were all computer literate and previous users of assistive technology, including the NVDA screen reader. Hence, the focus of the evaluation could turn to the functionality and quality of synthesised speech in the screen reader associated with the Qfrency TTS integration.

The majority of the respondents found it easy to navigate through a document or web page by using the keyboard. One respondent reported that the TTS does not indicate whether or not a question is followed by an empty line, nor that it gives auditory feedback on the pronunciation of letters as they are typed on the keyboard. This functionality had not yet been implemented in the Qfrency driver.

The end-users were generally satisfied with the quality of the Qfrency TTS voices. They commented favourably on the human-like features of the voices, including the breathing sounds of the English voice in particular, but also pointed out small technical issues. For instance, both the Afrikaans and English female voices sometimes had sharp sounds when pronouncing certain words, which hurt the ears. This can be attributed to signal processing artifacts that are difficult to control in the modelling process of the synthetic voices. It is an ongoing research topic.

The respondents had different requirements on certain aspects of the technology. One respondent indicated that the screen reader must be able to distinguish between words properly and adhere to punctuation when synthesising long academic texts. The rate of speech must be constant so that his concentration is not broken. He valued intelligibility higher than naturalness. In contrast, another respondent, who has serious hearing loss, indicated that he requires a voice which has a natural rhythm. He also commented that people with hearing loss tend to prefer male voices, as the lower frequency of male voices, compared to that of female voices, increases the user experience. Another respondent conceded that, if these voices were the only ones available to him, he would happily use them without complaining.

Overall, the respondents agreed that the Qfrency TTS voices could make a significant impact in the print-disabled community in South Africa, as the community previously had to make do with international (mostly English) voices to read documents or web pages. The respondents also appreciated the fact that they had been approached to conduct the evaluation, as they felt they had contributed to something that would change the way they interact with computers in the future.

4 USE CASE 3: ACCESSIBLE E-BOOKS

4.1 Background

The previous use case described the challenges that print-disabled and/or low-literate persons face when accessing text-based documents and web pages on a computer. We noted that a screen reader can assist these persons with navigation and reading of the content out loud by using TTS to generate audio dynamically. Another vehicle for accessible information delivery is a format of e-books that can embed audio (and other types of content such as image and video) along with the text, in a single file. The statically-produced audio can be human-narrated or synthesised. Daisy is an international standard for such accessible e-books that has been employed mostly by the print-disabled community. Its successor, EPUB 3, is a more versatile, inclusive standard that is being adopted by mainstream publishers of e-books [14].

However, current practice in South Africa is still to publish e-books either as text-only books or as audiobooks with limited navigation capabilities. In order to enhance the accessibility of these offerings, we engaged with select publishers of text-only books, and organisations that produce human-narrated audiobooks of mainstream works for the print-disabled community in the country. We integrated Qfrency TTS into an EPUB 3 conversion system to produce EPUB 3 e-books with synchronised text and audio, whether human-narrated or synthesised. This enables fine-grained search for blind end-users, as well as synchronised, highlighted reading for dyslexic end-users [9, 15]. The audiobook producers, as well as other organisations and individuals in the print-disabled community, evaluated the EPUB 3 conversion system and its output e-books.

4.2 System Integration

The EPUB 3 conversion system runs on Linux and has two main components. When human-narrated audio is available, the first component uses Qfrency TTS to align the text and audio in the EPUB 3 e-book at word level to enable fine-grained search for blind end-users, as well as synchronised, highlighted reading for dyslexic end-users. When only text is available, the second component uses Qfrency TTS to add synthesised audio to the EPUB 3 e-book, with implicit alignment for the search and synchronised reading functionalities, though currently only at sentence level due to integration challenges. Fig. 6 illustrates the conversion process.

For the first component, we utilised Aeneas, an open-source alignment tool developed by ReadBeyond [13]. Its alignment method is based on the dynamic time warping (DTW) signal processing technique. It synthesises the text of the EPUB 3 e-book by using a TTS voice in the language of the text and human-narrated audio. For this use case in our local languages, we integrated the Qfrency TTS voices into Aeneas through a custom driver. Aeneas then aligns the synthesised audio with the human-narrated audio by using DTW to match the similar content overlapping in the two audio signals and relate the timing information back to the text.

For the second component, we could leverage the Daisy pipeline, an accessible e-book format production and document conversion tool from the Daisy Consortium [3]. It is open-source and the latest version supports EPUB 3. One of its core functionalities is to use TTS to add synthesised audio to an e-book. It can run on multiple platforms and supports SAPI on Windows, but, because the rest of the bigger system only runs on Linux, we performed a custom integration of the Qfrency TTS voices into the Linux version.

In addition to an EPUB 3 validation tool, we used Readium, an extension to the Google Chrome browser [5], to test the synchronised text and audio in the output EPUB 3 e-books using the highlighting feature. Figs. 7 and 8 show screenshots of sample e-books being highlighted in Readium for English and isiZulu, respectively.

4.3 User Engagement

The evaluation process was two-fold and was conducted with two user groups. The system evaluators (as the first user group) had to test whether the two components of the EPUB 3 conversion system performed their functions correctly, namely aligning available human-narrated audio with text at word level, and generating synthesised audio from text, which is implicitly aligned at sentence level. They would execute the task remotely on the HLTRG Linux server using a text-based connection from their Windows desktop over the Internet. We gave instructions to the system evaluators on how to use the system. In addition, they had to comment on their experience while using the system, by filling in questionnaires. We refer to this task as the "system evaluation".

The end-user evaluators (as the second user group) had to test the acceptability of the output of the EPUB 3 conversion system, that is the EPUB 3 e-books with synchronised text and audio, both human-narrated and synthesised. The questionnaires included factors such as synchronisation (alignment) accuracy, quality of the TTS voice and general usability. We gave instructions to the evaluators on how to use Readium to read the e-books. Blind or partially-sighted evaluators could use the NVDA screen reader with Readium to complete the task. We refer to this task as the "end-user evaluation".

The sample e-books used in the two tasks covered the languages of South African English, Afrikaans, isiZulu, Sesotho, and Sepedi. Both human-narrated and synthesised audio were available in the first three languages, with only human-narrated audio in Sesotho and synthesised audio in Sepedi. Some respondents tested more than one language.

Three respondents, each from a different organisation, conducted the system evaluation of both types of EPUB 3 e-books. Their technical competence ranged from novice to expert. Everyone was able to generate the sample e-books successfully. All found it easy to do from the start, except one who struggled at first but then became used to how it works as he practised with more e-books. They opened these generated e-books in Readium and verified that the audio plays back and the synchronised highlighting works.

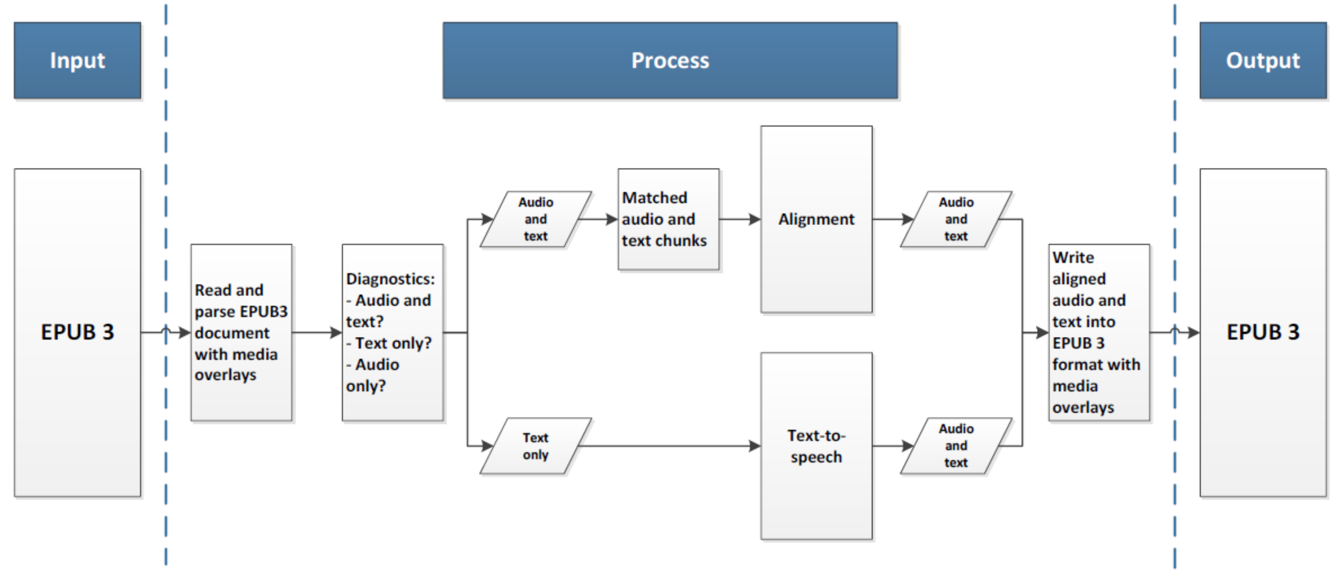
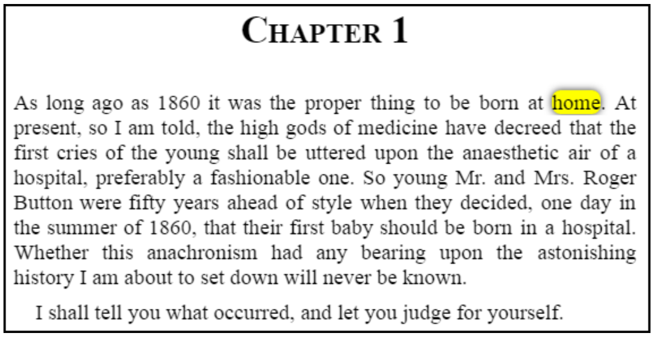

Input
Process
Output
EPUB 3
Read and parse EPUB3 document with media overlays
Diagnostics: - Audio and text? - Text only? - Audio only?
Audio and text
Matched audio and text chunks
Alignment
Audio and text
Text only
Text-to-speech
Audio and text
Write aligned audio and text into EPUB 3 format with media overlays
EPUB 3

Figure 6: EPUB 3 conversion system flow diagram

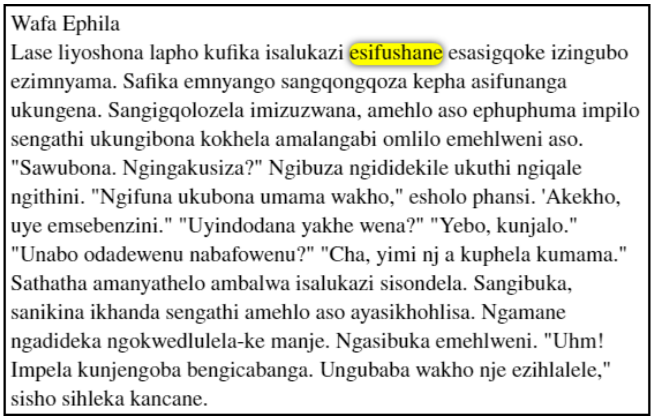


Chapter 1

As long ago as 1860 it was the proper thing to be born at home. At present, so I am told, the high gods of medicine have decreed that the first cries of the young shall be uttered upon the anaesthetic air of a hospital, preferably a fashionable one. So young Mr. and Mrs. Roger Button were fifty years ahead of style when they decided, one day in the summer of 1860, that their first baby should be born in a hospital. Whether this anachronism had any bearing upon the astonishing history I am about to set down will never be known.

I shall tell you what occurred, and let you judge for yourself.

Figure 7: Sample English EPUB 3 e-book highlighted in Readium



Wafa Ephila
Lase liyoshona lapho kufika isalukazi esifushane esasigqoke izingubo ezimnyama. Safika emnyango sangqongqoza kepha asifunanga ukungena. Sangigqolozela imizuzwana, amehlo aso ephuphuma impilo sengathi ukungibona kokhela amalangabi omlilo emehlweni aso. "Sawubona. Ngingakusiza?" Ngibuza ngididekile ukuthi ngiqale ngithini. "Ngifuna ukubona umama wakho," esholo phansi. 'Akekho, uye emsebenzini." "Uyindodana yakhe wena?" "Yebo, kunjalo." "Unabo odadewenu nabafowenu?" "Cha, yimi nj a kuphela kumama." Sathatha amanyathelo ambalwa isalukazi sisondela. Sangibuka, sanikina ikhanda sengathi amehlo aso ayasikhohlisa. Ngamane ngadideka ngokwedlulela-ke manje. Ngasibuka emehlweni. "Uhm! Impela kunjengoba bengicabanga. Ungubaba wakho nje ezihlalele," sisho sihleka kancane.

Figure 8: Sample isiZulu EPUB 3 e-book highlighted in Readium

The system evaluators were generally excited about the technology and what it can achieve. One even created an extra e-book from on-site content that was not part of the required samples. We asked the respondents and their colleagues whether they had any additional requirements regarding the user interface. People from the one organisation preferred a graphical user interface, similar to those of the audio editing applications they use in their audiobook production environment.

The respondents of the end-user evaluation, also across different organisations, included people with visual impairments, people working with the dyslexic community, people who have physical disabilities and people who did not have any form of disability. We did not require any technical expertise from them. Ten respondents evaluated the sample e-books with human-narrated audio and four evaluated the sample e-books with synthesised audio.

All the respondents were able to open the e-books with human-narrated audio in Readium, navigate the structure and play the audio with ease. They reported that the audio and highlighted words were aligned at normal playback speed. All but one found the same when increasing and decreasing the rate of playback. Everyone could perform the opening and playback of the e-books with synthesised audio easily, though the navigation took practice to master. The respondents confirmed the accuracy of the text and audio synchronisation at sentence level, even at variable speed, though one was concerned about the inhibitive length of sentences instead of words when trying to follow and read the highlighted text.

Most of the end-user evaluators had a positive experience with the EPUB 3 e-books. Across the human-narrated and synthesised audio groups, all but two respondents from the former group affirmed that they would read these e-books if they were the only versions of accessible e-books available. However, when testing the e-books with synthesised audio, some did perceive the Qfrency

TTS voices as monotonous. As noted in the first use case, research continues to improve the naturalness of the voices.

In terms of additional requirements, they would like to be able to search for a particular word or phrase in Readium, as that functionality is not yet implemented in the reader, even though the EPUB 3 e-books support it. With regard to technical issues, Readium exhibits some inconsistency in its highlighting behaviour on some Windows systems, as it selects the whitespace between words and sentences instead of the text itself. We also discovered a latency issue on some machines when using the Readium-NVDA combination that degraded the user experience. For both these issues, we suspect that the increased processing workload of the additional word and sentence-level markup in the e-books cause the instability. We will need to investigate whether performance enhancements must be made to the EPUB 3 e-book structure or to Readium.

5 CONCLUSION

The AAC use case is the first project on which the HLTRG and the CAAC collaborated to serve the South African AAC community. The specialist knowledge, skills, and networks of the CAAC enabled us to interact with AAC technology providers, therapists and end-users not only to determine the requirements of future multilingual AAC systems in the South African context but also to perform a baseline evaluation of the Qfrency TTS voices integrated into an AAC system. From a technical perspective, this integration was the cleanest, swiftest and easiest among all the use cases. It can be attributed to the simpler TTS requirements for AAC, the standardisation of SAPI and the mature product status of Grid 3. The favourable acceptance rate of the Qfrency TTS voices among the end-users confirmed the commercial status of the voices, albeit with room for improvement, as discussed. We note that an important contributing factor to the success of the evaluation was the training that the end-users had received from the CAAC on how to use AAC systems. The combined technological and sociological expertise of the HLTRG and the CAAC shows great potential to break down the communication barriers that these persons currently experience.

The screen reading use case proved more difficult to execute. There were many unforeseen integration challenges when we wanted to incorporate the Qfrency TTS voices into the NVDA screen reader. A screen reader requires more complex interactional features from a TTS engine to read paragraphs and other structure in documents and web pages, than what an AAC system requires to vocalise a single word, phrase or sentence in a cell. The results from the evaluation were incomplete due to the contrasting technology readiness levels of the end-users at the public ICT centres (low) compared to those at the private learning institution (high). This confirms the lesson from the AAC use case on how important it is to have a specialist training partner with an established end-user base during deployment. Nevertheless, the end-users had an overall positive impression of the Qfrency TTS voices and they were pragmatic in their opinion about the usability in the NVDA screen reader.

The use case for accessible e-books led to similar insights. The EPUB 3 standard is relatively new and, although numerous implementations exist in the e-book production and reading markets, many of these only cover certain aspects of the standard. The few that support the features that we require for synchronised text and audio still behave inconsistently. On the production side, one of the challenges was the ever-changing structure of the input EPUB 3 e-books to our conversion system that made automation difficult. On the reading side, the volatility and limited options of applications that support synchronised highlighting led to problems when testing the EPUB 3 e-books. The challenge to Qfrency TTS when it needs to align and synthesise the e-books is the same as in the screen reading use case, namely how to process document-level text more intelligently. The evaluators consisted of the audiobook producers and other organisations and individuals in the print-disabled community. Some provide specialist services to blind persons, hence their own participation and ability to facilitate the evaluation to other end-users were of immense value. We realise, however, that our understanding of the requirements of dyslexic readers is incomplete, and so we need to collaborate more formally with specialists in the dyslexic community in the future. The results from this first round of engagement show the promise of the accessible EPUB 3 e-books to improve the reading experience of print-disabled end-users. Future work will need to address the coverage and stabilisation of the EPUB 3 reading features in applications.

Finally, we recognise that the most suitable TTS voice for a particular use case will not necessarily be the same across the board. From the perspective of the AAC end-user, a baseline intelligible TTS voice might be a sufficient starting point, since, for the first time, she has "something" in her mother-tongue that is better than "nothing". On the other hand, her dialogue partner without a disability might set a higher standard of listening quality that, for example, includes more naturalness, before he would accept the synthesised speech. When a print-disabled individual uses a screen reader to navigate a computer or read academic material, she might insist on swift synthesis with a clear, understandable TTS voice. The TTS would then need to sacrifice naturalness for intelligibility to remain responsive enough. In contrast, when TTS is used to narrate leisurely material, its naturalness will become more important again. Though, this comes at a computational cost that could be inhibitive for dynamic use (produced in real time). Fortunately, EPUB 3 e-books with static synthesised audio (produced offline) can circumvent this problem. Future research, development, and application of the Qfrency TTS voices will need to remain cognisant of these factors.

6 ACKNOWLEDGEMENTS

We would like to thank the technology providers, therapists and other members of the AAC community for their participation in the first use case. The same gratitude goes out to the publishers, audiobook producers and other members of the print-disabled community for their involvement in the second and third use cases.

REFERENCES

[1] Olabode Akinsola, Marlien Herselman, and SJ Jacobs. 2005. ICT provision to disadvantaged urban communities: A study in South Africa and Nigeria. *International Journal of Education and Development using ICT* 1, 3 (2005).

[2] Ben Caldwell, Michael Cooper, L Guarino Reid, and Gregg Vanderheiden. 2008. Web content accessibility guidelines (WCAG) 2.0. *WWW Consortium (W3C)* (2008).

[3] Daisy Consortium. 2017. Daisy Pipeline 2. (2017). http://daisy.github.io/pipeline/

[4] J. Craig and M. Cooper. 2014. Accessible Rich Internet Applications 1.0. *W3C Recommendation* (2014).

[5] Readium Foundation. 2017. Readium extension for Google Chrome. (2017). http://readium.org/

[6] Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing* (second ed.). Pearson Education.

[7] G Kaisara and Shaun Pather. 2009. e-Government in South Africa: e-service quality access and adoption factors. (2009).

[8] Johannes A. Louw, Avashlin Moodley, and Avashna Govender. 2016. The Speect text-to-speech entry for the Blizzard Challenge 2016. In *Proceedings of The Blizzard Challenge 2016 Workshop*.

[9] Dominik Lukeš. 2015. Dyslexia friendly reader: Prototype, designs, and exploratory study. In *Information, Intelligence, Systems and Applications (IISA), 2015 6th International Conference on*. IEEE, 1–6.

[10] Sally Millar and Janet Scott. 1998. What is augmentative and alternative communication? An introduction. *Augmentative communication in practice: An introduction* (1998), 3–12.

[11] United Nations. 2008. Convention on the Rights of Persons with Disabilities (CRPD) Article 9 - Accessibility. (2008). https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html

[12] Department of Social Development. 2016. White Paper on the Rights of Persons with Disabilities. Government Gazette No. 39792. (2016). http://www.gpwonline.co.za/Gazettes/Gazettes/39792_9-3_SocialDev.pdf

[13] Alberto Petterin and contributors. 2017. Aeneas - a Python/C library and a set of tools for forced alignment. (2017). https://github.com/readbeyond/aeneas

[14] Sue Polanka. 2013. What librarians need to know about EPUB3. *Online Searcher* (2013), 70–72.

[15] Luz Rello, Horacio Saggion, and Ricardo Baeza-Yates. 2014. Keyword highlighting improves comprehension for people with dyslexia. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL*. 30–37.

[16] Paul Taylor. 2009. *Text-to-Speech Synthesis* (first ed.). Cambridge University Press.

[17] Smartbox Assistive Technology. 2017. Grid 3. (2017). https://thinksmartbox.com/product/grid-3/

[18] Kerstin M. Tönsing, Karin van Niekerk, Georg I. Schlünz, and Ilana Wilken. 2017. AAC services for multilingual populations: South African service provider perspectives. *Submitted for publication in a journal* (2017).

[19] Kerstin M. Tönsing, Karin van Niekerk, Georg I. Schlünz, and Ilana Wilken. 2018. AAC services for multilingual populations: South African end-user perspectives (working title). *To be submitted for publication in a journal* (2018).