Feasibility of individualised synthetic speech for children with complex communication needs in three South African languages (South African English, Afrikaans, and isiXhosa).

Thesis presented for the Degree of Doctor of Philosophy in the Division of Communication Sciences and Disorders, Faculty of Health Sciences, University of Cape Town

University of Cape Town

Name : Camryn Terblanche
Student No. : TRBCAM001
Supervisor : Michal Harty, University of Cape Town
Co-supervisor : Michelle Pascoe, University of Cape Town
Course Code : AHS7001W
Date : 26 July 2024

Plagiarism Declaration

I, Camryn Terblanche, hereby declare that the above thesis is my own unaided work, both in concept and execution, apart from the normal guidance from my supervisors and contributions from others as outlined in the acknowledgements.

I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own. I have used the American Psychological Association convention for citation and referencing where possible, unless specified otherwise by a specific journal. Each contribution to, and quotation in this PhD from the work(s) of other people has been attributed, cited, and referenced.

Neither the whole nor any part of the above thesis has been in the past, or is being, or is to be submitted for a degree at this university, or any other university.

I grant the University of Cape Town free licence to reproduce the above thesis, in whole or in part, for the purpose of research.

I am now presenting the thesis for examination for the degree of PhD.

Signature:

Student name: Camryn Terblanche

Student number: TRBCAM001

Date: 26 July 2024

Declaration on the Inclusion of Publications in a PhD Thesis

I confirm that I have been granted permission by the University of Cape Town's Doctoral Degrees Board to include the following publications in my PhD thesis, and where co-authorships are involved, my co-authors have agreed that I may include the publications:

Manuscript one: Phase 1a

- Terblanche, C., Harty, M., Pascoe, M., & Tucker, B. V. (2022). A situational analysis of current speech-synthesis systems for child voices: A scoping review of qualitative and quantitative. Evidence. *Applied Sciences*, *12*(5623), 1-17. https://doi.org/10.3390/app12115623

Manuscript two: Phase 1b

- Terblanche, C., Pascoe, M., & Harty, M. (2025). Challenges, perceptions and implications of AAC use in South African classrooms: An exploratory focus group study. *Child Language Teaching and Therapy*, *41*(1), 47–65. https://doi.org/10.1177/02656590241311063

Manuscript three: Phase 2

- Terblanche, C., Schnoor, T. T., Harty, M., & Tucker, B. V. (2024). The Development of Synthetic Child Speech in Three South African Languages. *Augmentative and Alternative Communication*. 1-12 https://doi.org/10.1080/07434618.2024.2374312

Manuscript four: Phase 3

- Terblanche, C., Pascoe, M., & Harty, M. (2025). Do you like my voice? Stakeholder perspectives about the acceptability of synthetic child voices in three South African languages. *International Journal of Language & Communication Disorders*, *60*(1), e13152. https://doi.org/10.1111/1460-6984.13152

Signature: Student number: TRBCAM001

Student name: Camryn Terblanche Date: 26 July 2024

Table of Contents

List of Tables

Supplementary Table 1. The Children's Responses, using a Pictographic 3-Point Scale, to Questions about the Quality, Acceptability, and Utility of the Synthetic Speech**.**

List of Figures

Abstract

Background:

A person's voice is an expression of their identity. The uniqueness of a person's voice is influenced by both their physical and social attributes. Yet, for children with complex communication needs (CCN), sometimes the only functional way to communicate is by using an augmentative and alternative communication (AAC) device, specifically speech-generating devices. However, AAC users often lack a personal connection to the synthetic voices found on speech-generating devices. While these devices improve the quality of life for those with speech impairments, they often fail to capture the unique linguistic diversity present in the population, along with the uniqueness of an individual's natural voice.

Aim and Objectives:

The overarching aim of this research is to develop a viable method for creating natural-sounding synthetic voices for South African children with CCN by using open-source speech synthesis software, taking into consideration the cultural assumptions and ideologies that influence the development of AAC systems for individuals with diverse backgrounds. The four primary objectives are 1) To outline the strengths and weaknesses of existing speech synthesis systems: This involves a detailed examination of the challenges and advancements in the speech synthesis field, 2) To document multiple stakeholders' experiences and perceptions: This objective aims to capture the challenges faced by professionals and caregivers when implementing AAC within schools for Learners with Special Education Needs (LSEN), and gather their ideas for overcoming these challenges, ensuring a comprehensive understanding of the context. 3) To delineate the process of generating naturalistic synthetic child speech: Tacotron 2, an open-source speech synthesis system, is used for three under-resourced languages (South African English/SAE, Afrikaans, and isiXhosa). This objective aims to provide insights into the feasibility of creating synthetic voices that match the vocal identity of children with CCN, addressing issues of speech diversity and under-resourced languages. 4) To evaluate and document multiple stakeholders' perspectives surrounding the quality, acceptability, and utility of newly created synthetic speech: This involves gathering feedback on the synthetic voices generated. This objective aims to ascertain whether the utilised method would be accepted and deemed appropriate as an addition to AAC in South Africa. Understanding stakeholder perspectives is crucial for refining synthetic voices and ensuring their practicality and acceptance in real-world contexts.

Methods:

The PhD project employs an exploratory, sequential, mixed method methodology. The PhD research comprises three distinct phases, Phase 1 begins with a scoping review (Phase 1a) which is followed by focus group discussions (Phase 1b) utilising a descriptive qualitative design. This initial exploration is

followed by Phase 2, where Tacotron 2 is employed for synthetic speech development. The assessment of the naturalness of the synthetic voices created during this phase follows a non-experimental, quantitative descriptive design. In the final phase (Phase 3), a mixed methods design, specifically a triangulation mixed method design, is adopted. This approach amalgamates qualitative insights gathered from focus groups with quantitative data, ensuring a comprehensive understanding of stakeholder perspectives and their broader assessment of the newly created synthetic speech.

Results:

The scoping review in Phase 1a uncovered several challenges in child speech synthesis, emphasising the need for tailored solutions considering the specific linguistic and age-related variations for children with CCN. In Phase 1b, AAC implementation challenges in South Africa revealed pervasive issues of reduced support, training, and crime-related safety concerns associated with the use of high-tech AAC devices. Limited accessibility further highlighted the barriers faced by children with CCN in LMICs. Phase 2's investigation into Tacotron 2's feasibility in generating synthetic child speech showed promising outcomes. Despite challenges like limited child speech data and literacy disparities among children providing the speech data, we were able to create synthetic voices in three under-resourced South African languages—SAE, Afrikaans, and isiXhosa, using Tacotron 2. In Phase 3, stakeholder perspectives on the quality and acceptability of newly created synthetic voices highlighted a generally positive response. Despite variations in prosody and intelligibility compared to natural child speech, stakeholders recognised potential benefits for children with CCN, with intelligibility ratings averaging 92%. The synthesis of qualitative and quantitative data enriched the understanding of the synthetic voices' practicality and acceptance, contributing to future AAC solutions for children with CCN in South Africa and similar contexts.

Conclusions:

Collectively, this PhD research provides holistic insights into child speech synthesis, AAC implementation challenges, and stakeholder perspectives, especially in LMICs. The implications for service provision, safety, language diversity, and stakeholder involvement are evident. The findings lay the groundwork for advancing AAC interventions, promoting accessibility, and fostering inclusive decision-making processes, thereby enhancing communication solutions for children with CCN.

Acknowledgements

Thank you to the National Research Foundation, Mitacs, and The University of Cape Town for providing financial support, which enabled me to complete my PhD and travel to Canada and Prague. Due to these trips, I was able to make huge progress in my research and formulate wonderful partnerships and international networks. I am so thankful for the incredible opportunities offered to me during these last few years.

Thank you to my supervisors and collaborators, Michal Harty, Ben Tucker, and Michelle Pascoe. Your support, encouragement, and guidance have been incredible. This has been such a special time in my life, and you were a big part of that. You have taught me so much. A special thank you to Tyler Schnoor, for helping me with the coding. You made my idea possible, and I will be forever grateful.

Thank you to each one of my participants. You have been so amazing to work with. Thank you to each school for welcoming me with open arms. Thank you especially to the three learners who these voices were made for. Your genuine excitement, love, patience, and ability to persevere despite all the challenges you face is both inspiring and humbling. I would do this 100 times over, just to see your little faces light up again. Thank you for the impact you've had on my life.

I am beyond grateful to the Angel Network and iStore South Africa for donating the iPads and iPad covers for the three children. Thank you to Investec for connecting us with these generous donors. Gifting the iPads with the loaded speech to the children was an incredibly memorable and rewarding experience. Special thanks to Dayna Tate and Candice Waldeck for beautifully capturing the children's expressions as they received their iPads.

I couldn't have done this without my family, especially my mom and my sister. Thank you for celebrating each small milestone with me, and for picking me up when there was a delay. I wouldn't have been able to do this without your support, love, and belief in me (and all the prayers!).

"Take delight in the Lord, and he will give you the desires of your heart."

- Psalms 37:4

Style, Abbreviations and Key Terms

A note on spelling and style convention: UK English spelling and APA referencing has been used throughout this thesis, aside from one of the manuscripts titled *The development of synthetic child speech in three South African languages,* where US English spelling was used as it was published in a US journal. Due to specific journal requirements, terms may be used interchangeably throughout this thesis, for e.g., Speech-Language Therapist vs Speech-Language Pathologist and Complex Communication Needs vs Expressive Communication Difficulties.

Abbreviations

AAC: Augmentative and Alternative Communication

ASD: Autism Spectrum Disorder

CCN: Complex Communication Needs

CMLLR: Constrained Maximum Likelihood Linear Regression

CP: Cerebral Palsy

CSMAPLR: Constrained Structural Maximum A Posteriori Linear Regression

DNN: Deep Neural Networks

GAN: Generative Adversarial Network

GMMs: Gaussian Mixture Models

HMMs: Hidden-Markov-Models

LMICs: Low- and Middle-Income Countries

LSEN: Learners with Special Education Needs

LVCSR: Large-Vocabulary Continuous-Speech-Recognition

MAP: Maximum A Posteriori

MCCs: Mel-Cepstral Coefficients

MFCCs: Mel-Frequency Cepstral Coefficients

MLLR: Maximum Likelihood Linear Regression

MOS: Mean Opinion Scores

NAE: North American English

OTs: Occupational Therapists

PhD: Doctor of Philosophy

PLP: Perceptual-Linear-Prediction

SA: South Africa

SAE: South African English

SAT: Speaker-Adaptive Training

SGD: Speech-Generating Device

SLAM: The School Aged Language Assessment Measures

SLPs: Speech-Language Pathologists

SLTs: Speech-Language Therapists

TTS: Text-to-Speech

UK: United Kingdom

US: United States

USA: United States of America

VOCA: Voice-Output Communication Aid

VTLN: Vocal Tract Length Normalisation

WER: Word Error Rate

WHO: World Health Organisation

Key Terms

1. **Augmentative and alternative communication:** AAC includes methods and tools used to support or replace speech for individuals with communication difficulties. AAC can take many forms. No-tech or low-tech approaches include things like gestures, facial expressions, writing, drawing, spelling words by pointing to letters, or using pictures and symbols. High-tech AAC might involve using communication apps on tablets or speech-generating devices that produce voice output. Often, a person will use a mix of these tools, depending on the situation. The combination of methods someone uses is referred to as their AAC system (American Speech-Language-Hearing Association, 2022).
2. **Child speech synthesis:** Child speech synthesis involves creating synthetic voices that mimic the speech patterns of children, often using advanced speech synthesis technologies to achieve naturalness and intelligibility (Jain et al., 2022).
3. **Complex communication needs:** Children living with CCN, also known as expressive communication difficulties, may exhibit either unintelligible speech or limited verbal abilities. These challenges can arise from developmental conditions such as cerebral palsy or Autism Spectrum Disorder, or stem from acquired disorders such as traumatic brain injuries or stroke (Da Fonte & Boesch, 2019).
4. **Intelligibility:** Intelligibility is the degree to which speech can be understood by listeners. It measures how clearly and accurately synthetic speech conveys the intended message (Jette et al., 2017).
5. **Learners with special education needs:** LSEN refers to students who require specialised support and adaptations in their educational environment due to various disabilities or learning challenges (South African Department of Education, 2001).
6. **Low- and middle-income countries:** LMICs are countries with a gross national income per capita below a certain threshold, typically characterised by lower levels of development and limited resources compared to high-income countries (Hamadeh et al., 2023).

7. **Mean opinion scores:** In this context, MOS is a subjective measure used to evaluate the quality of synthetic speech by collecting listeners' opinions on various attributes such as naturalness and understandability. For instance, a 5-point Likert scale rating system is used to measure the quality of the synthetic voices, where 1 means "horrible" and 5 means "best" (Sefara et al., 2019).
8. **Quality:** In the context of speech synthesis, quality refers to the overall effectiveness of the synthetic speech in terms of clarity and listener satisfaction (Sefara et al., 2019).
9. **Speech naturalness:** Naturalness refers to how closely synthetic speech resembles natural human speech in terms of prosody, tone, and expressiveness (Sefara et al., 2019).
10. **Speech-generating devices:** SGDs are electronic devices that generate spoken output through symbol selection or text input. They can provide a means of communication for individuals who cannot rely on their natural speech to communicate effectively (Drager, Reichle, et al., 2010).
11. **Stakeholder perspectives:** Stakeholder perspectives refer to the viewpoints and feedback from individuals or groups who have an interest or role in the subject of the research, such as AAC users, caregivers, and professionals (Tönsing et al., 2018).
12. **Tacotron 2:** Tacotron 2 is a neural network-based text-to-speech system that uses deep learning to generate high-quality speech from text, incorporating attributes like speaker and language with ease of adaptation to new data. (Wang et al., 2017).
13. **Under-resourced languages:** Languages are considered under-resourced when they have a limited online presence, there is a lack of linguistic expertise on the language, there are limited data for speech and language processing, reduced transcribed speech corpora and pronunciation dictionaries, as well as limited resources for speech, language, and literacy development (Besacier et al., 2014).
14. **Warm start:** Warm start refers to initialising a system or process with pre-existing knowledge or data to improve performance and reduce time required compared to starting from scratch (Phuong et al., 2021).

1. Introduction

1.1. General Introduction

In a world where our voice is a key to our personal identity, the absence of a unique voice for children with complex communication needs (CCN), highlights a significant gap in assistive technologies for people with disabilities. When individuals cannot rely on their natural speech for communication, they often explore alternative communication methods (Jette et al., 2017). Augmentative and alternative communication (AAC) provides a vital solution for those whose speech does not meet their communication needs (Light et al., 2019). Children living with CCN, who experience expressive communication difficulties, may exhibit either unintelligible speech or limited verbal abilities. These challenges can arise from developmental conditions such as cerebral palsy or Autism Spectrum Disorder, or stem from acquired disorders such as traumatic brain injuries or stroke (Drager, Light, et al., 2010). Effective communication is crucial to learning, yet children with CCN often face barriers and exclusion in the classroom due to inadequate communication tools (Dada et al., 2016). This can result in diminished intentional communication, limited engagement with language and literacy, and fewer opportunities for social interaction (Drager, Light, et al., 2010), often resulting in communication partners underestimating their potential. The impact of these communication barriers extends beyond education, affecting access to appropriate healthcare, family dynamics, community involvement, and future employment prospects (Tönsing & Dada, 2016). To address these challenges, AAC technologies have become essential tools for children with CCN, enabling communication with unfamiliar communication partners.

Understanding the different types of AAC systems is crucial, as it sets the stage for exploring how speech synthesis technology can be applied in these systems. The classification of AAC systems encompasses several key dimensions, including the level of technology (low-technology/low-tech vs. high-tech), the type of output (non-verbal, digitised, or synthetic speech), the display type (static vs. dynamic), and the symbol format (graphic vs. text-based) (Beukelman & Light, 2020). Low-tech AAC include paper-based boards, while high-tech AAC includes speech-generating devices/ SGDs, capable of producing spoken output that is either digitised (recordings of natural speech) or synthesised (generated in real time through text-to-speech technologies), through symbol selection or text input (American Speech-Language-Hearing Association, 2022). One of the key differentiators in AAC systems is the type of display. Static displays offer a fixed set of options on a single page or board, commonly seen in low-tech systems, although they can also be incorporated into digital platforms. In contrast, dynamic displays are interactive, enabling AAC users to navigate across multiple pages or categories. This flexibility allows AAC users to construct more complex and nuanced messages by

selecting symbols that lead to new screens with related vocabulary. Dynamic displays are especially beneficial in high-tech systems, particularly for AAC users who require access to extensive vocabulary sets or multiple languages. AAC systems also vary in the types of symbols they utilise. These symbols can range from graphic representations, such as line drawings, photographs, or pictograms, which are often suited to individuals with limited literacy skills, to text-based systems for those who are literate (Beukelman & Light, 2020). Many AAC systems incorporate both types of symbols, offering versatile communication options for users with diverse needs. Understanding how these components work together is essential when selecting or developing AAC tools, as each configuration offers distinct benefits depending on the user's language abilities, motor skills, cognitive profile, and cultural context.

These AAC technologies, particularly those involving speech synthesis, have had a transformative impact, as demonstrated by the renowned physicist Professor Stephen Hawking, who relied on a speech-generating device in his daily life. Using such devices enhances the quality of life for individuals with speech impairments (Creer et al., 2013). A personalised synthetic voice on a speech-generating device allows young individuals to express their thoughts and connect with others, helping them to navigate the world with a voice that reflects their unique identities and communication needs, whilst alleviating feelings of social isolation (Jreige et al., 2009).

While communication technology is advancing, the communication technology specifically designed for children is lagging behind, despite its significance on a child's developmental trajectory. Instances of a child using an adult voice or multiple children sharing the same voice in a classroom setting highlight the importance of aligning the synthesised voice with the AAC user's identity (Mills et al., 2014). An individual's voice serves as a distinctive marker encompassing various attributes, including physical size, age, gender, race, intellectual ability, geographical and social background, as well as personality (Jreige et al., 2009; Mills et al., 2014; Sutton et al., 2019). However, Mills et al. (2014) highlighted a limitation in commercially available devices, emphasising that the speech output often fails to reflect important aspects such as the AAC user's age, sex, and personality. Furthermore, the languages provided on these commercially-available devices do not adequately represent the linguistic diversity found in countries such as South Africa, including their various distinctive dialects (Sefara et al., 2019). Using pre-loaded speech outputs in unfamiliar languages or AAC users' lacking a personal connection with the device's voice has the potential to cause embarrassment, decrease social motivation and may hinder the effective use of the AAC devices (Jreige et al., 2009; Tönsing et al., 2018). Conversely, implementing a personalised voice has the potential to reduce AAC abandonment, as it authentically represents the AAC user, creating a meaningful connection between them and the generated synthetic voice. However, the mismatch between available synthetic voices and the diverse identities of AAC users emphasises the importance for more inclusive and personalised solutions.

In addition, AAC support has traditionally centred on individuals with CCN, often overlooking the needs of caregivers, service providers, peers, and other communication partners. However, it is widely acknowledged that the success of AAC communication hinges not only on the AAC user but also on the collaboration with their communication partners (Creer et al., 2013; Jette et al., 2017; Kent-Walsh et al., 2015; Moorcroft et al., 2019b). The significance of early AAC intervention for children with CCN cannot be overstated, as it lays the foundation for communicative success. The efficacy of intervention is tied to caregiver acceptance and the active modelling of AAC, particularly in the context of young children (Moorcroft et al., 2019b). When children with CCN receive support from stakeholders such as caregivers, teachers, and therapists who embrace and encourage AAC usage, children with CCN are more likely to integrate their AAC systems into regular communication, offering them a means to comprehend others and express themselves in a manner that can be easily understood. This, in turn, contributes to enhanced participation in the classroom environment (Moorcroft et al., 2019b). Stakeholder acceptance is particularly important in a country such as South Africa. To offer additional background information, the South African context will be described.

1.2. The South African Context

South Africa categorises as an upper middle-income country according to the world bank (Hamadeh et al., 2023) and stands as a diverse, vibrant, and uniquely multicultural nation, boasting twelve official languages, including South African Sign Language (SASL), alongside numerous unofficial languages. In the Western Cape in 2022, Afrikaans (41.2%), isiXhosa (31.4%), and English (22%) emerged as the predominant languages spoken in households (Statistics South Africa, 2022). Low-and-middle-income countries (LMICs) are typically defined by a gross national income per capita below a certain threshold and are often characterised by lower levels of infrastructure, limited technological access, and constrained healthcare and educational resources when compared to high-income countries (Hamadeh et al., 2023). These challenges are particularly evident in the field of augmentative and alternative communication (AAC), where access to culturally and linguistically appropriate technologies remains limited. While speech synthesis systems are readily available for major languages like English and European languages, they remain limited for South Africa's indigenous languages, failing to adequately represent the country's rich linguistic and social diversity (Sefara et al., 2019). Compounding this, the prevalence of US-accented speech, influenced by market dynamics, further diminishes the authenticity of representation on commercially available speech-generating devices and AAC applications (Yamagishi et al., 2012).

In the educational landscape, South Africa's Education White Paper 6 (South African Department of Education, 2001) affirms the right of children with disabilities, including those with CCN, to an inclusive school environment. However, the country is progressing slowly towards the realisation of

this goal. Despite the flexibility granted by South Africa's official language policy, which permits schools to choose any of the official languages for teaching and learning, English often takes precedence over local African languages (South African Government, 1997). Essentially, South African children, including those with CCN, frequently find themselves learning and communicating in their second or third language within the school environment. Moreover, the under-served status of indigenous language speakers is exacerbated by the scarcity of contextually relevant resources in African languages, perpetuating a significant language mismatch often experienced between speech and language therapists (SLTs) and their clients (Pascoe & Norman, 2011). Therefore, challenges arise when striving for equality within the context of multilingualism, multiculturalism, and various societal issues, as highlighted by Pascoe and Norman (2011). In South Africa, it is reported that as many as 70% of school-age children with disabilities do not participate in formal education. Among those who are enrolled, the majority are often placed in segregated special schools designed for Learners with Special Education Needs (LSEN) (Donohue & Bornman, 2014). Regrettably, within these environments, there is inconsistent assurance of access to resources, the availability of rehabilitation personnel, and the provision of suitable assistive technology (van Niekerk et al., 2019). In addition, roughly 63% of children in South Africa are believed to come from impoverished households (Statistics SA, 2020). Within this demographic, many confront heightened levels of crime, overcrowded living conditions, generational illiteracy, and elevated instances of trauma and violence (van Niekerk et al., 2019). The challenges posed by these environmental factors substantially hinder children's access to education. As a result, a considerable percentage of South African children from economically disadvantaged backgrounds struggle to meet the exit requirements from secondary school, consequently impeding their transition into tertiary educational environments. It is clear that South Africa faces challenges in its education system, with disparities in language representation, a slow move towards inclusive environments, and difficulties in providing resources for learners with disabilities. However, within these challenges lies an opportunity for positive change. By recognising these issues, there is potential for innovative solutions that prioritise inclusive education, ensuring that every child, including those with CCN, has an equal chance for a brighter future.

1.3. Research Questions

The preceding section has shed light on the intricate South African landscape. The diverse population and varying income levels in South Africa create a complex yet dynamic environment. In the field of speech synthesis, this complexity is heightened by the mismatch between the synthetic voices on commercially-available speech-generating devices and the diverse identities of AAC users in LSEN classrooms, which emphasises the necessity for personalised solutions. This PhD research therefore seeks to comprehensively explore child speech synthesis technology within a LMIC context, gaining

insights into the challenges and opportunities associated with AAC implementation. The research is guided by four pivotal research questions outlined below.

1. What speech synthesis systems are currently available for child voices, and what are the gaps in the existing literature?
2. What are professionals' and caregivers' perspectives of AAC implementation within a LSEN educational setting, and what ideas do they propose for overcoming associated challenges?
3. What is the process of generating naturalistic synthetic child speech, matching the vocal identity of three children with CCN, using Tacotron 2, an open-source speech synthesis software, for three under-resourced South African languages?
4. What are the perspectives of multiple stakeholders regarding the quality, acceptability, and utility of newly created synthetic speech in three under-resourced South African languages?

In summary, the research questions collectively contribute to a holistic understanding of the AAC landscape in South Africa and explore the potential of utilising open-source software to develop feasible solutions to challenges faced by researchers and clinicians in under-resourced settings. Addressing these questions empowers researchers and practitioners to develop targeted interventions, policies, and resources to enhance communication accessibility for individuals with CCN in low-and middle-income countries, such as South Africa, who often experience resource constrained settings.

1.4. Outline of the PhD Research Project

In article format, this PhD research project followed a sequential structure organised into three distinct phases, each designed to address specific objectives and involving different participants and methodologies. Phase 1 was further subdivided into two stages, Phase 1a and Phase 1b. Subsequently, Phase 2 and Phase 3 were sequentially implemented in the research framework. In Phase 1a, the initial manuscript, a scoping review, explores the landscape of global speech synthesis systems for child voices, examining 58 studies spanning from 2006 to 2021. Focused on children aged between 2–16 years old, the review explores languages, methods, speech data, intelligibility, and age-related aspects of current speech synthesis systems. The primary goal was to identify existing speech synthesis systems for child voices and provide a comprehensive overview of the current technological landscape. Through this exploration, the review highlights available resources while discerning gaps in the literature, paving the way for a more nuanced understanding of the global state of speech synthesis systems for children.

In Phase 1b, the second manuscript describes stakeholders' perspectives of the current state of AAC implementation and use in LMICs, particularly South Africa, and highlights the pivotal role communication partners play in the successful implementation of AAC. This manuscript considers the perspectives of professionals (teachers and SLTs) and caregivers of children with CCN regarding AAC

implementation in under-resourced LSEN schools and highlights unique challenges encountered in such contexts. While some challenges align with those seen in high-income countries, the study reveals additional context-specific issues relating to AAC implementation in LMICs, including concerns about crime, affordability of high-tech AAC devices, device features, and accent variations. These findings were used to ensure that our subsequent AAC developments were not only functional but also realistic and beneficial for the South African population. Moreover, by gathering insights and potential solutions from stakeholders, this research contributes some practical strategies for overcoming barriers to effective AAC use in these contexts as described by these stakeholders.

In Phase 2, the third manuscript investigates the process of generating naturalistic synthetic child speech using Tacotron 2, an open-source speech synthesis software, for three under-resourced South African languages, namely South African English/SAE, Afrikaans, and isiXhosa. With a scarcity of child speech corpora, speech data from three neurotypically developing 11- to 12-year-old children were used to create synthetic child speech. Adult models were used to "warm start" the child speech synthesis process. Instead of training the model from scratch using only the child speech data, the researchers initially utilised adult models that had been pretrained on a broader dataset. This initialisation with adult models provided a starting point for the child speech synthesis, allowing the model to inherit some knowledge and patterns learned from the adult data. The subsequent fine-tuning process focused on adapting the model to the specific characteristics of child speech, aiming to match the vocal identity of three children with CCN, using the limited available child speech data from the neurotypically developing children. The research set out to successfully produce synthesised child voices of adequate quality in each language, as rated by 124 listeners. Phase 2 specifically addresses the challenge of developing synthetic voices that align with the age, social, and linguistic background of children with CCN in various under-resourced languages.

In Phase 3, the final manuscript examines the perspectives of multiple stakeholders, including children with CCN, professionals (teachers and SLTs), and caregivers of children with CCN regarding the quality, acceptability, and utility of the synthetic speech generated in Phase 2. Stakeholder feedback, including assessments of quality, intelligibility, and naturalness of the synthetic child voices, is crucial to understand the practical implications of implementing synthetic speech and ascertaining whether the utilised synthesis method is accepted and deemed appropriate as an addition to AAC in South Africa. Phase 3 will emphasise the importance of locally relevant voices for diverse AAC users.

In conclusion, this PhD project conducted an exploration of the complexity of providing synthetic speech output to children who need AAC and who are living in LMICs, specifically focusing on LSEN contexts in South Africa. Each of the three distinct phases serve a specific purpose, with Phase 1 mapping the global landscape of speech synthesis systems for child voices, to recognising the AAC

implementation challenges in LMICs. The subsequent phases address the identified gaps from Phase 1, with Phase 2 leveraging recent technological advancements to demonstrate the feasibility of creating synthetic child speech in under-resourced languages. In Phase 3, this novel product is coupled with the perspectives of stakeholders, including children with CCN, professionals, and caregivers, to determine the utility of the created voices. Together, the findings from this project illuminate the practical implications of implementing synthetic voices in these contexts, and emphasise the importance of creating tailored, culturally relevant voices for children from diverse backgrounds.

2. Methodology

2.1. Research Aim

The aim of this PhD research was to develop an inclusive, comprehensive and feasible approach for generating natural-sounding and personalised synthetic voices designed for South African children with CCN using SAE, Afrikaans, and isiXhosa. This was achieved through the utilisation of readily available open-source speech synthesis software, taking into consideration the cultural assumptions and ideologies that influence the development of AAC systems for individuals with diverse backgrounds. There were four research objectives:

1. To outline the strengths and weaknesses of existing speech synthesis systems, achieved by summarising the evidence base concerning the development of synthesised speech for children.
2. To document multiple stakeholders' experiences, perceptions of implementing AAC in their context, and their ideas for overcoming the challenges surrounding the use of AAC within LSEN schools in South Africa.
3. To delineate the process of generating naturalistic synthetic child speech, matching the vocal identity of three children with CCN, using open-source speech synthesis software, Tacotron 2, for three under-resourced South African languages, namely SAE, Afrikaans and isiXhosa.
4. To evaluate and document multiple stakeholders' perspectives surrounding the quality, acceptability, and utility of newly created synthetic speech in three under-resourced South African languages.

Table 1 gives an outline of the overall project methodology, as well as the methods, objectives and participants for each of the three PhD project phases.

Table 1. Summary of the methodology used in each of the project phases

Overarching research question Given the AAC landscape in South Africa, is it feasible to generate individualised, natural-sounding synthetic child speech in different under-resourced languages, and if so, what are the stakeholders' perspectives of the synthetic speech created?				
Study phase	Phase 1		Phase 2	Phase 3
	Phase 1a	Phase 1b		
Rationale	To identify available systems for child voices and pinpoint gaps in the literature through a scoping review of existing literature. This involves a	To understand the AAC needs and challenges of stakeholders in the community, and their ideas for overcoming these challenges. By	First, to gather spontaneous speech samples from children for the purpose of crafting an individualised synthetic voice; second, to establish a	To evaluate the intelligibility and naturalness of the synthetic voices, aiming to ascertain whether the utilised method

	comprehensive examination of the current state of knowledge and available resources in the field.	doing so, we can try to ensure that the developments in AAC are not only functional but also realistic and beneficial for the SA population.	comprehensive and feasible method for researchers to create synthetic voices.	would be accepted and deemed appropriate as an addition to AAC in South Africa.
Objectives	To outline the strengths and weaknesses of existing speech synthesis systems, achieved by summarising the evidence base concerning the development of synthesised speech for children.	To document multiple stakeholders' experiences, perceptions of implementing AAC in their context, and their ideas for overcoming the challenges surrounding the use of AAC within LSEN schools in South Africa.	To delineate the process of generating naturalistic synthetic child speech, matching the vocal identity of three children with CCN, using open-source speech synthesis software, Tacotron 2, for three under-resourced South African languages, namely SAE, Afrikaans and isiXhosa.	To evaluate and document multiple stakeholders' perspectives surrounding the quality, acceptability, and utility of newly created synthetic speech in three under-resourced South African languages.
Participants	No participants	• Teachers • SLTs • Caregivers	• SAE speaking child with CCN • SAE speaking neurotypically developing child • Afrikaans speaking child with CCN • Afrikaans speaking neurotypically developing child • isiXhosa speaking child with CCN • isiXhosa speaking neurotypically developing child • 124 adult listeners	• Teachers • SLTs • Caregivers • 3x children with CCN
Design	Exploratory, sequential, mixed method methodology			

	Scoping review	Descriptive qualitative	Non-experimental quantitative descriptive	Triangulation mixed method
Method	Scoping review	Focus group discussion	Collection of speech data, creation of synthetic speech, MOS ratings	Focus group discussion, MOS and WER ratings
Analysis	Arksey and O'Malley framework stages for conducting scoping reviews combined with the Levac *et al.* (2010) enhancements.	Reflexive thematic analysis	Descriptive and inferential statistics	Reflexive thematic analysis, descriptive and inferential statistics
Manuscripts	Manuscript 1: A situational analysis of current speech-synthesis systems for child voices: A scoping review of qualitative and quantitative.	Manuscript 2: Challenges, perceptions, and implications of AAC use in South African classrooms: An exploratory focus group study.	Manuscript 3: The Development of Synthetic Child Speech in Three South African Languages.	Manuscript 4: Do you like my voice? Stakeholder perspectives about the acceptability of synthetic child voices in three South African languages.

2.2. Conceptual Framework

2.2.1. The social justice theory and social inclusion

Guided by the foundational principles of justice and fairness, social justice theory, as articulated by Rawls (1971), asserts that all individuals should enjoy equal rights and liberties, and no person should be deprived of opportunities available to others in society. The ultimate outcome of just and fair treatment for all, regardless of differences, is social inclusion. Acknowledging the profound impact of social inclusion on the well-being of children with CCN and their families, South Africa, in accordance with the guidelines set forth in Education White Paper 6 (South African Department of Education, 2001), asserts the right of children with disabilities, including those with CCN, to experience an unrestricted and inclusive school environment. In the pursuit of social justice and inclusive education, the transformative shift from a dualistic (special and general) to an inclusive education system reflects a dedication to creating educational spaces that prioritise equal opportunities, diversity, and the overall holistic development of every child, regardless of their communication needs or linguistic background. With this project, children with CCN are not only accommodated but also empowered through appropriate AAC tools. This is particularly important in a multilingual context like South Africa, where language plays a significant role in identity and belonging.

Moreover, the World Health Organization (WHO) distinguishes "disability" from "impairment", highlighting that individuals may have impairments, but it is systemic barriers, negative attitudes, and exclusion from society, whether intentional or unintentional, that contribute to disabling people (Anastasiou & Kauffman, 2013). This perspective highlights the transformative potential of inclusive practices in preventing impairments from inevitably resulting in disabilities. Social justice and inclusion stand as the foundational principles guiding this PhD research project, where everyone, regardless of their abilities, their background, or the language they speak, is included and empowered. In this endeavour, AAC emerges as a critical tool, facilitating effective communication for individuals with diverse social and linguistic needs. Acknowledging the significance of AAC within the broader framework of inclusive education, it becomes an instrumental means to dismantle barriers, empower individuals, and cultivate a truly inclusive learning environment where every unique voice, irrespective of communication abilities, is valued and heard. This PhD research project ensures that all children, regardless of the language they speak, have access to high-quality speech synthesis for their AAC device.

2.2.2. Culturally responsive theoretical framework

According to Amery et al. (2022), the past two decades have witnessed a growing body of global AAC research. However, the current literature reveals a persistent influence of Western culture on both the AAC research processes and outputs (Amery et al., 2022; Tönsing et al., 2019). To acknowledge the cultural assumptions and ideologies that shape AAC system development for individuals from diverse backgrounds, including those from LMICs, this PhD research positions itself within a culturally responsive theoretical framework, adapted from Kirkhart and Hopson (2010). This framework is structured around five types of validity: methodological, interpersonal, theoretical, experiential, and consequential validity. Emphasising the recognition of various ways of knowing and meaning-making rooted in culture and context, this framework asserts that multiple perspectives should be respected, particularly those from historically marginalised groups. This includes AAC stakeholders from LMICs and children with CCN, whose voices are often marginalised in society. Table 2 shows the researcher's culturally responsive research approach, with examples of how each dimension was considered in the PhD project.

Table 2. Culturally responsive research approach, with examples of how each dimension was considered in the current project

Dimension of validity	**Considerations in the current project**
Methodological validity "The cultural appropriateness of measurement tools and cultural congruence of design	- Participant selection is representative of diverse linguistic, social, and cultural backgrounds. - The materials used considered the diverse backgrounds of the participants (including literacy levels of the children).

configurations"(Kirkhart & Hopson, 2010, p. 13).	- The languages used to create the synthetic speech reflected the languages most spoken by the Western Cape population and matched the linguistic background of the three children with CCN.
Interpersonal validity "The quality of the interactions between and among participants in the evaluation process" Kirkhart & Hopson, 2010, p. 13).	- Researchers reflected on their own cultural positions and positions of authority. - Researchers ensured confidential and anonymous data collection processes. - Focus groups were separated to ensure power imbalances did not occur between professionals and caregivers.
Theoretical validity "The cultural congruence of theoretical perspectives underlying the program, the evaluation, and assumptions of validity" Kirkhart & Hopson, 2010, p. 13).	- The synthetic voices created reflect the diverse linguistic and cultural backgrounds of South African children, as well as their personal preferences, thus supporting social inclusion and addressing the needs of marginalised children. - Feedback was integrated from stakeholders from diverse cultural groups and grounded in real-world cultural experiences. - The synthetic speech technology was accessible to three children from under-resourced backgrounds, making it equitable, regardless of their socioeconomic or linguistic status.
Experiential validity "Congruence with the life experience of participants in the program and in the evaluation process" Kirkhart & Hopson, 2010, p. 13).	- Researchers incorporate stakeholders' perspectives in the analysis process. - Researchers ensure diverse perspectives are represented in the data. - Research data are understood in terms of the realities of the people they represent (i.e., stakeholders from LMIC).
Consequential validity "The social consequences of understandings and judgments and the actions taken based upon them" Kirkhart & Hopson, 2010, p. 13).	- Researchers made every effort to accommodate diverse literacy levels and language preferences when interacting with stakeholders, ensuring that communication was accessible and inclusive. - Researchers disseminate findings in a way that takes community partners' voices into account. This included conducting member checking after focus groups to validate interpretations and ensuring that participant confidentiality and anonymity were maintained. Recommendations derived from stakeholders' feedback are therefore relevant and actionable for the community.

2.3. Research Design

This PhD research project follows an exploratory, sequential, mixed methods design. A mixed-methods design is defined as "research in which the investigator collects and analyses data, integrates the findings and draws inferences using both qualitative and quantitative approaches or methods in a single study" (Tashakkori & Creswell, 2007, p. 4). This project is structured into three distinct phases, each tailored to specific objectives and involving different participants and methods. This phased approach ensures that the study addresses its research questions effectively by employing suitable methodologies aligned with the respective objectives in each phase. Figure 1 gives an illustration of the research design.

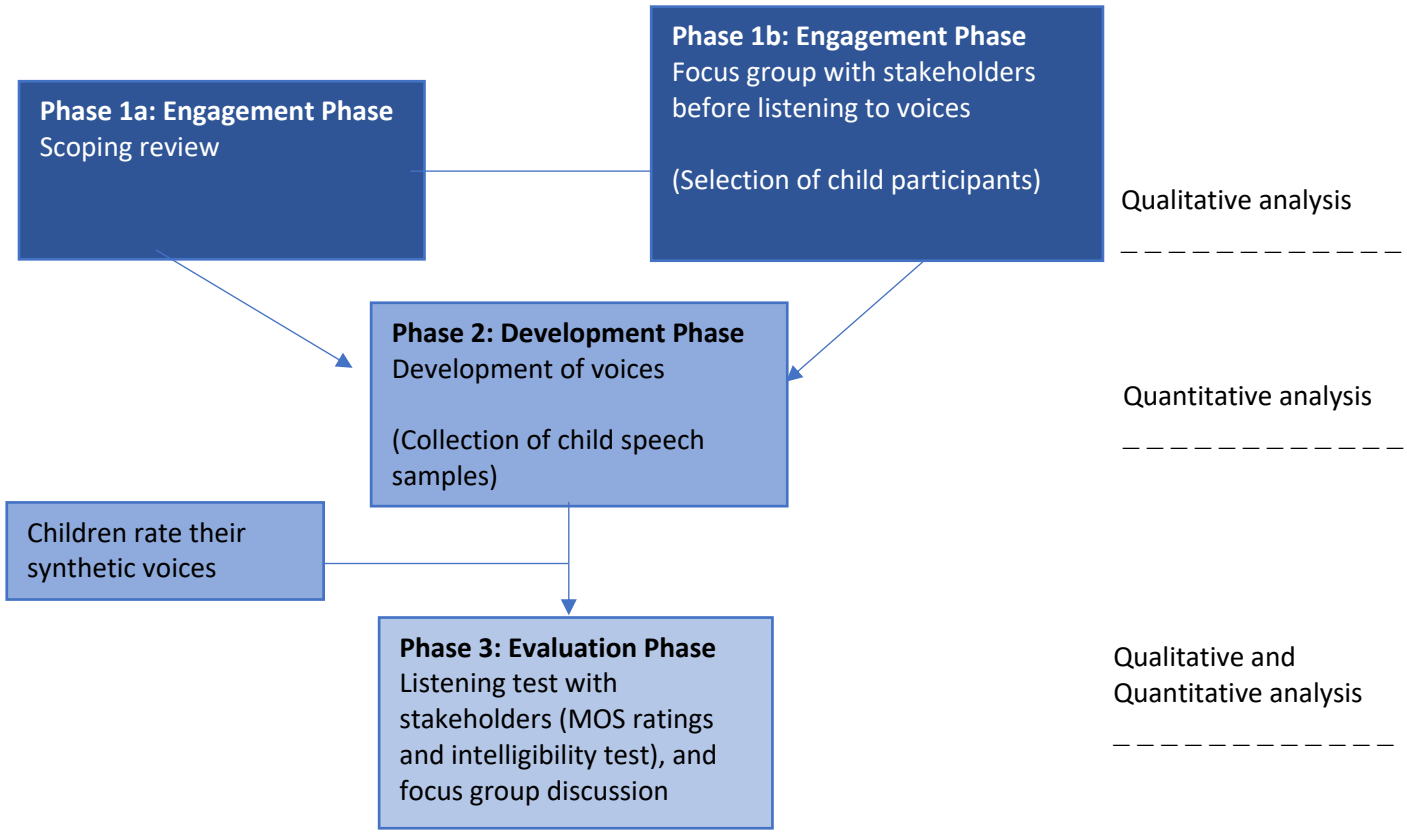


Figure 1. An illustration of the overall research design.

The study's three phases follow a carefully designed sequential structure. Phase 1 (Engagement phase) is further divided into two stages: Phase 1a involves a scoping review, and Phase 1b consists of a descriptive qualitative design using focus group discussions. During Phase 1b, a descriptive qualitative design was employed to provide detailed descriptions from various stakeholders (Terre blanche et al., 2006). In Phase 2 (Developmental phase), the development of synthetic voices is exploratory, while the assessment of their naturalness adopts a non-experimental quantitative descriptive design. Phase 3 (Evaluation phase) utilises a mixed methods design, specifically a triangulation mixed method design, allowing for the expansion of quantitative results with qualitative data (Creswell & Clark, 2007). In this phase, focus groups with stakeholders facilitated the collection of both qualitative and quantitative data.

Reflexive thematic analysis is used to analyse data from the focus group discussions (Braun & Clarke, 2006, 2019, 2021), while inferential and descriptive statistics aid in interpreting intelligibility tests and mean opinion scores. This comprehensive approach ensures a thorough analysis of the research objectives at each phase of the study.

2.4. Participants

In this study, the researcher started by first identifying and selecting the three children with CCN who were deemed most suitable for the development of synthetic voices. The criteria for selection included attendance at a LSEN school in Cape Town and proficiency in one of three different home languages: SAE, Afrikaans, or isiXhosa. These three languages were chosen as they are the three official provincial languages in the Western Cape (Western Cape Language Committee, 2020), where the research took place. The selection criteria specified children between the ages of 9;0-13;0 years old, diagnosed with CCN, and either candidates for AAC or already utilising an AAC system for communication. Essential skills included the ability to engage in a task, comprehend that a symbol or line drawing represented a word or concept, select a picture from a field of at least four, and demonstrate some residual speech abilities. Additionally, the children needed to employ direct selection methods, such as finger pointing. Exclusion criteria comprised children operating at a concrete object level, inability to recognise pictures due to visual or cognitive impairment, severely limited fine-motor skills, significant auditory processing problems, or those lacking any residual speech abilities. An assessment of these capabilities was conducted through a short screener administered by the researcher.

After the selection of the three children with CCN, the identification of other participants commenced. Figure 2 provides an overview of the participants involved in different phases of the study, with Phase 1a involving no participants as it constituted a scoping review. For Phase 1b, participants included teachers, SLTs, and caregivers. Most of these participants were acquainted with the three previously identified children with CCN. The SLTs and teachers were required to be involved in teaching, working, or consulting with children with CCN. Caregivers were family members and/or guardians of the three children with CCN (parents, grandparents, aunts, uncles, etc.). The researcher aimed to involve at least four participants per child, including caregivers, teachers, and SLTs.

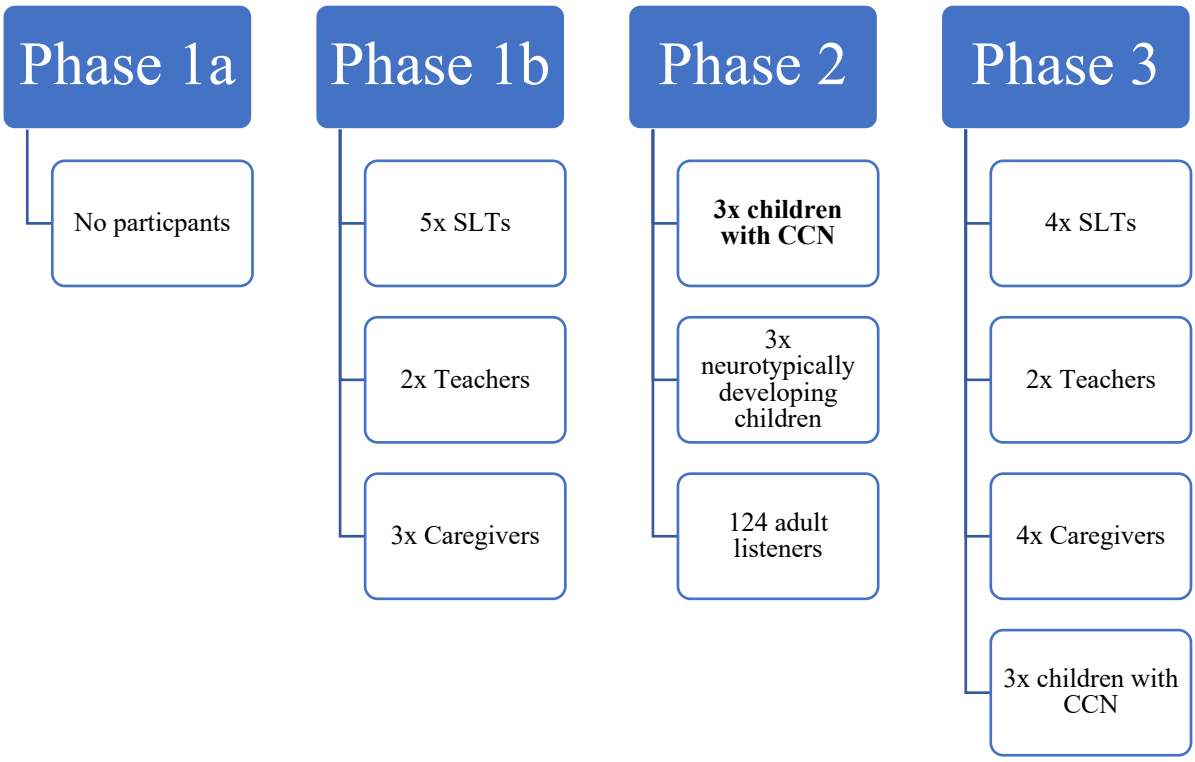


Figure 2. An overview of the participants included in the study.

In Phase 2, spontaneous (5 min) and read speech (5 min) samples from 98 neurotypically developing children were obtained at one mainstream English school in Cape Town. The selection of these neurotypically developing children was based on their similarity to the three children with CCN, for whom the synthetic child speech was intended. Similarities included factors such as age, gender, home language, and demographic group. Although the children with CCN were selected first, it became evident after conducting Phase 1a that the development of synthetic voices would require speech samples from neurotypical children whose vocal characteristics closely resembled those of the children with CCN. This understanding informed the age criteria from the outset. Children between the ages of 9 and 13 years were considered developmentally suitable for this purpose, as neurotypical children typically demonstrate greater speech intelligibility, fewer articulation and reading errors, and sufficient literacy skills to read in their home language. These abilities were important for producing acoustically clear and phonetically accurate speech samples, which are essential for high-quality synthesis.

With that in mind, the selection of the neurotypical children considered their performance in the picture description and reading tasks. From the pool of 98 neurotypically developing children, only three were chosen to further participate in the study due to their close resemblance to the target characteristics of the children with CCN, and their literacy skills. It's important to note that although the children with CCN were selected first, their speech samples were only collected in Phase 2 of the study.

In Phase 2, synthetic speech was generated using the speech data collected from both neurotypically developing children and children with CCN. Once the synthetic speech had been created, adult participants were invited to participate in a survey aimed at rating the naturalness of the synthetic speech. The survey included questions about the participants' spoken languages, and if they did not speak a particular language/s, the relevant questions related to synthetic voices in that language did not appear. Participation was open to any South African individual over the age of 18.

In Phase 3, participants who were involved in Phase 1b were invited to return for a second focus group session. Additionally, the three children with CCN were asked to provide their opinions about the voices that had been created specifically for them. This phase aimed to gather valuable insights and feedback from both adult participants and the children with CCN, whom the voices were created for.

2.4.1. Methods of recruitment

Permission was granted by UCT, Faculty of Health Sciences Human Research Ethics committee (HREC ref: 765/2021) and the Western Cape Education Department. Following this, various SLTs were contacted. Figure 3 gives a summary of the recruitment process used.

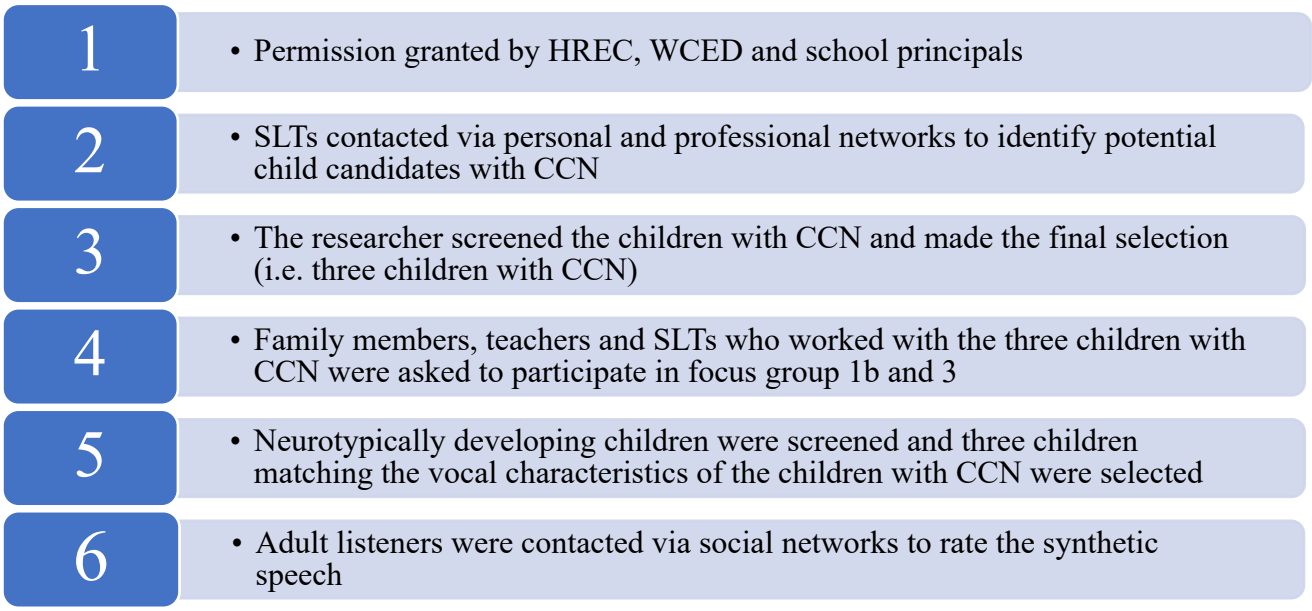


Figure 3. Summary of recruitment process

A) Recruitment of children with CCN (Purposive sampling)

1) Numerous SLTs at various schools were contacted via personal and professional networks to identify potential child candidates with CCN that fit the selection criteria (e.g., Vera School for Autistic Learners, Blouvlei School, Molenbeek School, Agape School, Astra School,

Glenbridge School, Eros School, Bel Porto School, Tembalethu LSEN school, Beacon school for LSEN and Vista Nova).

2) The principals gave permission to access the schools.
3) Caregivers were sent an information letter (in their home language) and asked to sign a consent form, giving permission for their children to be screened.
4) The researcher screened the children selected by the SLTs and made the final selection (three children, one in each language). The children with the most residual and intelligible speech were given preference. The final three children attended three different LSEN schools in Cape Town.
5) The three children's parents/legal guardians were asked to sign a final consent form, giving permission for their children to participate in the study.
6) The three children with CCN gave assent through picture selection.

B) Recruitment of caregivers (Purposive sampling)

Once the three children with CCN were identified:

1) Two of the children's caregivers (parents, legal guardians, grandparents, siblings etc.) were asked to participate in the focus groups (i.e., one focus group *before* they heard the synthetic voice [Phase 1b] and another focus group *after* they heard the synthetic voice [Phase 3]).
2) The caregivers were provided with an information letter and a consent form.

C) Recruitment of teachers and SLTs (Purposive sampling)

1) The principals gave permission to access the schools.
2) The children's teachers (primary educator, assistants etc.) and therapists (SLTs, occupational therapists etc.) were asked to participate in two focus groups [Phase 1b and Phase 3].
3) The professionals were provided with an information letter and a consent form.

D) Recruitment of neurotypically developing children (Purposive sampling)

1) The principal gave permission to access the mainstream school.
2) All caregivers were given an information letter and a consent form for their children to be audio recorded and screened. Each child gave assent to be screened before continuing.
3) Based on the neurotypically developing children's vocal similarity to the three children with CCN, as well as their performance in the picture description and reading tasks, three children were recruited.
4) The caregivers of these three children gave further consent for their children to participate in the study.
5) Each of these three children were told about the study and gave assent before continuing.

E) Recruitment of adult listeners (Snowball sampling)

1) Adult listeners in Phase 2 were asked to take part in an online survey. They were given an information sheet about the study, and they gave consent before participating in the survey.

2) Colleagues and friends were approached to give their opinion on the voices and invited to share the link with others. The survey was shared via anonymous link over social media and email.

2.5. Materials and equipment

Table 3 and 4 provides a description of the materials and equipment used in this study.

Table 3. Description of materials

Materials	**Description of materials**
Test of Aided-Communication Symbol Performance (TASP) (Bruno, 2010)	The TASP is a criterion referenced AAC assessment tool which can be adapted for use in any language. It was used as a screening tool for the selection of the child participants with CCN. It can be used within a wide variety of clinical populations, including children with autism spectrum disorder, apraxia, CP, and developmental delay. The TASP is a paper tool, which is helpful for individuals who initially find it difficult to stay on task when technology is introduced. The TASP can be used with children who have the ability to point, and it takes 10-20 minutes to complete. It includes four subtests: 1) symbol size and number, 2) grammatical encoding, 3) categorisation and 4) syntactic performance.
The School Aged Language Assessment Measures [SLAM])	The SLAM was used as a screening tool in the mainstream school. Firstly, the neurotypically developing children looked at a wordless picture story- The Ball Mystery Story, which is designed for children aged between approximately 9-15 years old. The children were asked to tell a story in their home language (English, Afrikaans and/or isiXhosa), based on the pictures. Their responses were audio recorded. Following this, each child was asked a number of questions about the pictures. SLAM was developed by Crowley (2021) to assess comprehension (understanding the questions posed) and expressive language. Furthermore, one can learn about a child's narrative skills, syntax, inferences/problem solving skills, cohesion, theory of mind, perspective taking, social/pragmatic language and dynamic learning in school aged children.
Multilingual library books	Age-appropriate story books, in each child's preferred language, were used.

Focus group interview schedules	Four focus group interview schedules were created. 1. Phase 1b caregiver group 2. Phase 1b professional's group (SLTs and teachers) 3. Phase 3 caregiver group 4. Phase 3 professional's group (SLTs and teachers)
Afrikaans (Louw & Schlünz, 2016a), isiXhosa, (Louw & Schlünz, 2016c) and the SAE Lwazi III text-to-speech datasets (Louw & Schlünz, 2016b)	Each Lwazi III dataset is made up of short and long audio clips from one female adult speaker per language. Each language has approximately 6-7 hours of data.
Tacotron 2 model (tacotron2_statedict.pt) from NVIDIA	A published pre-trained model, trained on the LJ dataset (Ito & Johnson, 2017) was used. The LJ dataset (Ito & Johnson, 2017) is extensive, with approximately 24 hours of short recordings from one female adult North American English (NAE) speaker (Ito & Johnson, 2017).
Clips of synthetic speech	Various short clips of synthetic child and adult speech (created during Phase 2) were used in the evaluation phase of Phase 2 as well as Phase 3 of the study.
Speech synthesis evaluation sheets	In Phase 3, two evaluation sheets were required. 1. AAC-based evaluation sheet for children with CCN 2. MOS and intelligibility evaluation sheet for caregivers and professionals

Table 4. Description of equipment

Software and/or equipment	**Description of software and/or equipment**
A Zoom H1 Handy recorder (44100 Hz, 16 bit),	A portable digital voice recorder was used to record focus group discussions and collect the speech samples from the neurotypically developing children, and the children with CCN.
Nvivo	A computer software for qualitative data analysis.
Tacotron 2	Tacotron 2 is a state-of-the-art, open-source speech synthesis system, which generates synthesised speech directly from graphemes and consists of a recurrent sequence-to-sequence mel-spectrogram prediction network (Wang et al., 2017).

Titan V GPU	A powerful computer, which was used to run the speech synthesis models.
jsPsych	jsPsych is a JavaScript framework for creating behavioural experiments that run in a web browser. This was used to run the survey in Phase 2.

2.6. Data Collection Procedures

Careful and confidential data collection procedures were followed throughout this project. Figure 4 visually outlines the data collection process across each phase of the study, providing a clear overview of how data was gathered and organised throughout the project.

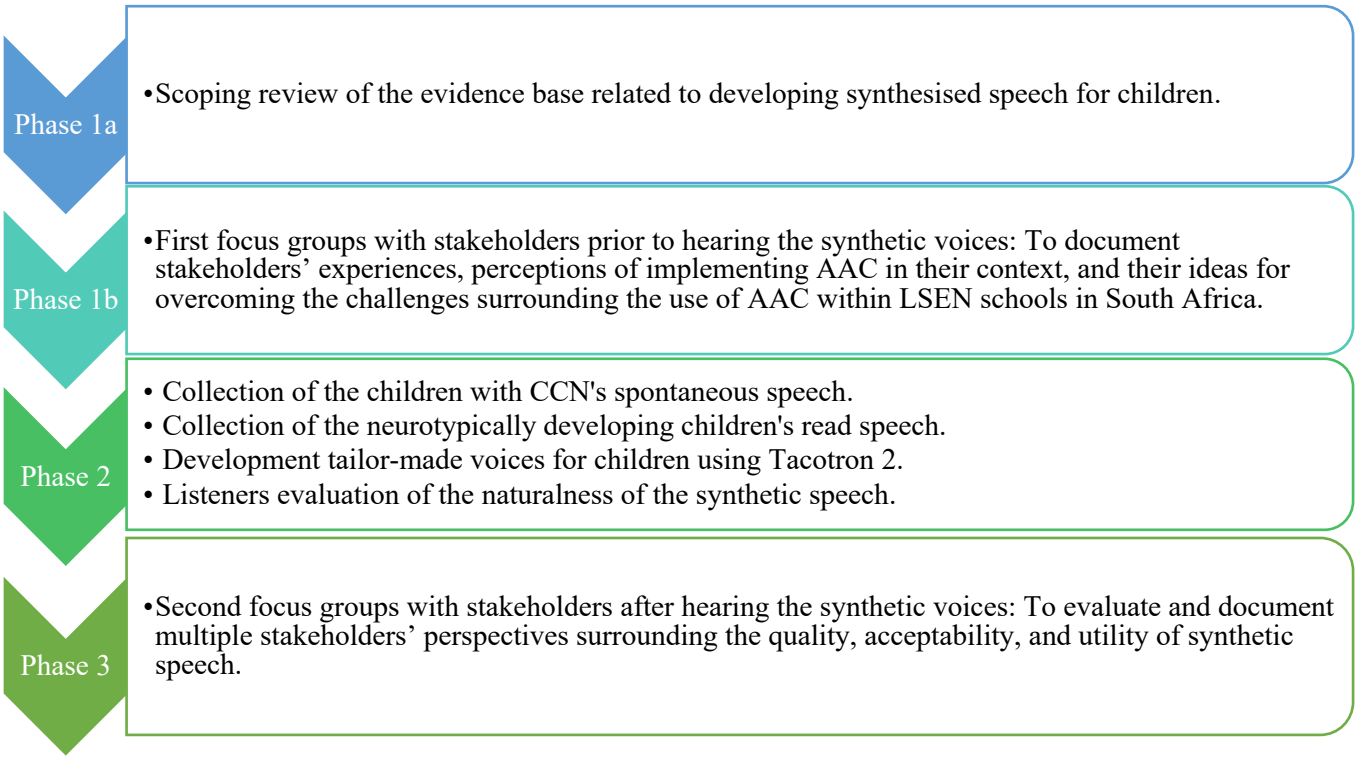


Figure 4. Visual representation of data collection per phase of the study

Following the scoping review in Phase 1a, Phase 1b aimed to explore stakeholder experiences, perceptions of the barriers and facilitators to implementing AAC in their context, and their ideas for overcoming the challenges associated with AAC. This exploration was conducted through two separate focus groups—one for caregivers (with a minimum of 2 participants per child, totalling at least 6 participants) and another for professionals (including at least two teachers and/or SLTs per child, totalling at least 6 participants). The division into separate focus groups ensured that caregivers' voices were distinctly heard without unintentional influence from professionals. Participants were offered translators during the focus groups to facilitate communication.

In Phase 2, spontaneous and read speech samples were collected from the children. Subsequently, the synthesis procedure commenced, leading to the creation of synthetic voices. To evaluate the naturalness of these synthetic voices, adult listeners participated in an online subjective mean opinion score (MOS) task. MOS is a performance metric employed to gauge the quality of speech based on subjective evaluations (Sefara *et al.*, 2019).

As with Phase 1b, there were two focus groups in Phase 3 (caregivers and professionals). Phase 3 provided stakeholders who participated in the initial Phase 1b focus groups with the opportunity to review the newly created synthetic voices. To collect quantitative data, stakeholders assessed the performance and quality of the synthetic voices using an in person subjective MOS task. Additionally, stakeholders were asked to transcribe sentences, allowing the researcher to evaluate intelligibility through word error rates (WER). Intelligibility focuses on individuals' ability to comprehend the synthesised speech (Sefara et al., 2019). Participants then took part in a focus group discussion to ascertain whether the utilised method would be accepted and deemed appropriate as an addition to AAC in South Africa.

2.7. Data Analysis

Phase 1a:

As per Colquhoun et al. (2014), we followed the Arksey and O'Malley framework stages for the conduct of scoping reviews combined with the Levac *et al.* (2010) enhancements when conducting the scoping review. Figure 5 shows the steps followed during the scoping review. A composite search strategy was executed to ensure all studies meeting the selection criteria were identified, whilst avoiding a biased evidence base. The search strategies were drafted by the PhD researcher and one of her supervisors, who had expertise in electronic search strategies and further refined through team discussion.

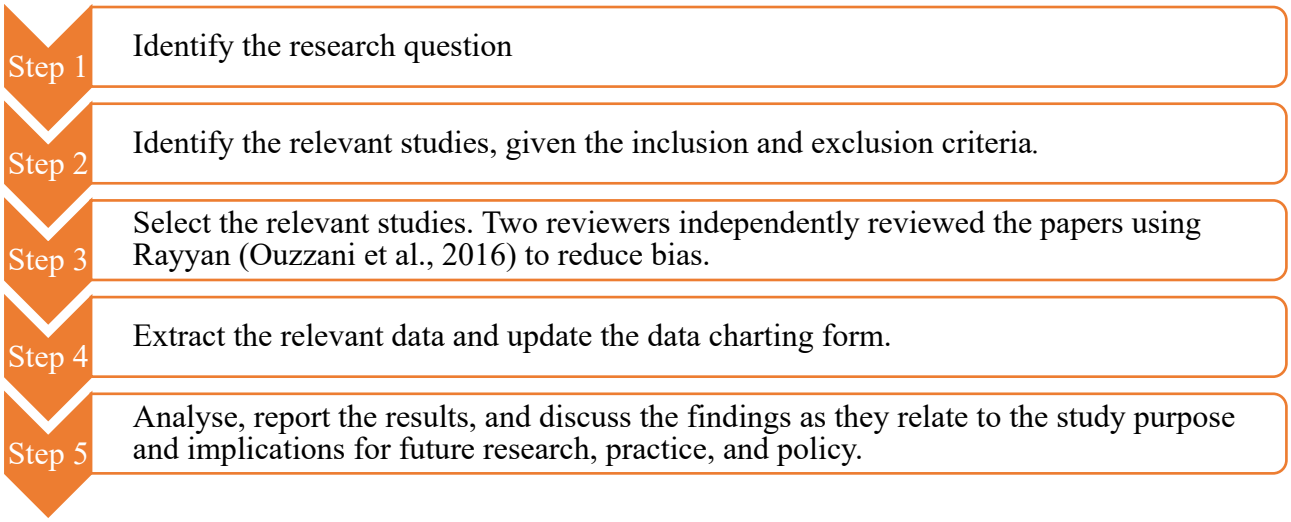


Figure 5. The steps to conducting a scoping review (Colquhoun et al., 2014)

Phase 1b:

Verbatim transcripts of the focus group discussions were analysed via reflexive thematic analysis, as originally outlined in 2006 (Braun & Clarke, 2006, 2019, 2021), facilitated by NVivo (QSR International, 1998). Reflexive thematic analysis is a theoretically flexible interpretive method for qualitative data analysis, enabling the identification and examination of patterns or themes within a dataset (Braun & Clarke, 2006, 2019, 2021). Two raters independently applied the coding framework to the focus group data. Discrepancies were discussed until consensus was reached and 100% agreement between the two raters occurred, ensuring validity and reliability of the analysis and the identification of key information.

Phase 2:

In the online survey, each listener listened to semantically predictable sentences (SPS), as per Govender and de Wet (2016), and made 18 MOS ratings per language: 5 MOS ratings for the synthetic child speech implementing warm start A, 5 MOS ratings for the synthetic child speech implementing warm start B, and 8 MOS ratings for the synthetic adult speech. Using R (R Core Team, 2019), both descriptive and inferential statistics were conducted on the MOS data.

Phase 3

As mentioned previously, stakeholders were asked to evaluate the performance and quality of the synthetic voices using an in-person subjective MOS task. Participants rated the overall impression, pleasantness, naturalness, understandability, and similarity to speakers, using a 5-point Likert Scale. The participants were also asked to judge the intelligibility in a sentence comprehension task. The participants were asked to listen to the synthetic speech samples and transcribe the audio. Semantically predictable sentences, as suggested by Govender and de Wet (2016), were used, and the average WER was calculated for each synthesis system. Dependent variables included the % correct responses on the sentence comprehension task and the ratings of overall impression of the system, pleasantness, naturalness, understandability, and similarity. Using R (R Core Team, 2019), both descriptive and inferential statistics were conducted on the MOS data. Verbatim transcripts of the focus group discussions were analysed via reflexive thematic analysis, incorporating Braun and Clarke's (2006, 2019, 2021) five-stage framework approach, and using NVivo (QSR International, 1998).

2.8. Ethical Considerations

Permission to conduct the study was obtained from the Human Research Ethics Committee of the University of Cape Town Health Sciences Faculty (HREC Reference number: 765/2021). Permission to access participants was obtained from the Western Cape Education Department and the relevant school principals. Ethical considerations for this study were guided by the Helsinki Declaration of 2013 (World Medical Association, 2013). These include:

2.8.1. Autonomy

All participants were invited to voluntarily participate in the study. They were explicitly informed of their right to withdraw at any time without facing negative consequences, aligning with ethical considerations (Leedy & Ormrod, 2013). Adult participants provided informed consent before participating. For child participants, written consent was sought from parents or legal guardians, and the child participants were also asked to give assent through picture selection. This approach ensures that both legal guardians and children were actively involved in the decision-making process regarding participation.

2.8.2. Beneficence

Beneficence, as per ethical considerations in research, entails maximising the benefits afforded to participants (Terre blanche et al., 2006). In this study, child participants with CCN were allowed to retain their synthetic voices after the completion of the study, as three iPads were donated to the children. Additionally, parents and teachers of the neurotypically developing children in the study received feedback following the picture description task, serving as a helpful literacy and language screener for teachers, and benefiting the children in an academic setting. While the study results may not be universally applicable to the entire population of child AAC users, the process used to create synthetic speech was thoroughly described. The intention is that this contribution serves as a model which can be replicated by others interested in developing synthetic child voices, particularly in under-resourced languages spoken in low- and middle-income countries globally.

2.8.3. Non-maleficence

Non-maleficence was prioritised through the implementation of safety measures in all study procedures. For instance, child speech samples were collected in secure environments, namely their schools, and in the presence of family members or school personnel, ensuring a supportive context (World Medical Association, 2013). Breaks were provided when the children felt tired, and the researcher ensured that

they did not miss essential activities such as break time, special school events, or important academic lessons. Furthermore, the researcher was mindful of the venues for the focus groups, ensuring that participants, especially caregivers, could easily access them without bearing any associated travel costs. Confidentiality was rigorously maintained through the use of assigning each participant a unique code, which guaranteed the anonymity of both data and participant information. The codes were stored separately from the list of participants, reinforcing the commitment to safeguarding the confidentiality and privacy of all involved.

2.8.4. Justice

The principle of justice, emphasising fairness within specific population groups, was upheld in this study (Terre blanche et al., 2006). Participants were selected based on their suitability, availability, and willingness to take part, ensuring a fair and respectful approach. Regardless of their social or economic backgrounds, all participants were treated with respect and fairness throughout the study. To further enhance fairness, the study offered translators and provided forms in each participant's home language, ensuring that language differences did not pose a barrier to participation. This approach aimed to promote inclusivity and equitable access to the study, aligning with the principle of justice.

3. Manuscripts

MANUSCRIPT ONE

A situational analysis of current speech-synthesis systems for child voices: A scoping review of qualitative and quantitative evidence.

Camryn Terblanche [1], Michal Harty [1], Michelle Pascoe [1] and Benjamin V. Tucker [2]

[1]Divsion of Communication Sciences and Disorders, University of Cape Town, Cape Town 7700, South Africa; trbcam001@myuct.ac.za (C.T.); michal.harty@uct.ac.za (M.H.); michelle.pascoe@uct.ac.za (M.P.)
[2]Department of Linguistics, University of Alberta, Edmonton, AB T6G 2R3, Canada; benjamin.tucker@ualberta.ca

Abstract

(1) *Background:* Speech synthesis has customarily focused on adult speech, but with the rapid development of speech-synthesis technology, it is now possible to create child voices with a limited amount of child-speech data. This scoping review summarises the evidence base related to developing synthesised speech for children. (2) *Method:* The included studies were those that were (1) published between 2006 and 2021 and (2) included child participants or voices of children aged between 2–16 years old. (3) *Results:* 58 studies were identified. They were discussed based on the languages used, the speech-synthesis systems and/or methods used, the speech data used, the intelligibility of the speech and the ages of the voices. Based on the reviewed studies, relative to adult-speech synthesis, developing child-speech synthesis is notably more challenging. Child speech often presents with acoustic variability and articulatory errors. To account for this, researchers have most often attempted to adapt adult-speech models, using a variety of different adaptation techniques. (4) *Conclusions:* Adapting adult speech has proven successful in child-speech synthesis. It appears that the resulting quality can be improved by training a large amount of pre-selected speech data, aided by a neural-network classifier, to better match the children's speech. We encourage future research surrounding individualised synthetic speech for children with CCN, with special attention to children who make use of low-resource languages.

Keywords: augmentative and alternative communication (AAC); children; complex communication needs; neural networks; speech synthesis

Introduction

Children with complex communication needs (CCN) may have developmental conditions such as autism-spectrum disorder, cerebral palsy or Down syndrome, or they may have an acquired disorder as a result of a traumatic brain injury or stroke, which results in disordered speech abilities. Children with CCN fall within a spectrum, as some present without intelligible speech whilst others have developed minimal speech but cannot primarily rely on their speech to communicate (Drager, Light, et al., 2010). When someone cannot rely on their own speech to communicate, they often look to other techniques to provide them with an alternative means to communicate. One such alternative is to use augmentative and alternative communication (AAC) devices. An AAC device with the capability for speech synthesis is referred to as a speech-generating device (SGD) or a voice-output communication aid (VOCA). The use of SGDs has been shown to increase the quality of life for individuals with speech impairments (Creer et al., 2013).

Without the ability to communicate, individuals may withdraw from social interaction, and even from interaction with their own family. This is often compounded when they are required to use a synthetic voice that sounds robotic, sounds like someone from a different geographical or social background, or someone of a different age or sex (Yamagishi et al., 2012). In fact, it is not uncommon to see a nine-year-old girl using an adult male voice or several children in a classroom using the same voice (Begnum et al., 2012; Mills et al., 2014). Speech-synthesis technology has traditionally focused on adult speech. Developing synthetic child speech could be considered more challenging for researchers, and even more so for children with CCN who make use of low-resource languages. Despite these challenges, researchers have begun exploring new techniques and methods for natural and intelligible-sounding child-speech synthesis.

Communication is essential to learning; without access to functional communication, children with CCN are often restricted from participating in the classroom (Dada et al., 2016). This may result in fewer instances of intentional communication, minimal language and literacy development, and poor opportunities for socialisation (Drager, Light, et al., 2010). As a result, their potential is often underestimated. This has repercussions for other areas of their life, such as accessing healthcare, participating in family and community activities, as well as engaging in future employment (Tönsing & Dada, 2016). Due to the importance of communication in our social, educational, professional, and personal lives, it is essential that the speech output of SGDs be as intelligible as possible. Using natural speech as the benchmark, *intelligibility* is defined as the accuracy with which an acoustic signal is conveyed by a speaker and recovered by a listener (Jette et al., 2017). A synthesised voice with high intelligibility is thus required for the acceptability of the SGD, not only in terms of it providing the service for which it was designed, but also in terms of the positive attitudes towards the voice and the

social interaction with the user (Creer, 2009). In contrast to intelligibility, in which the speech signal is extracted from the context, *comprehensibility or* understandability is the degree to which speech is understood when combined with available relevant information (i.e., linguistic context or conversational topics) (Jette et al., 2017). Speech naturalness and pleasantness could be considered important components of comprehensibility.

Speech naturalness can be described as how well the speech matches a listener's standards of rate, rhythm, intonation, and stress. Therefore, speech naturalness determines how natural the speech sounds to the listener and is compared to the speech they regularly hear in their immediate social environment (Sefara et al., 2019). Using a voice with a low level of naturalness, such as a robotic voice, does not match the human that it represents. Creer (2009) suggested that people interact differently with machines, and if the perception is that they are addressing the communication aid rather than the human user, their interactional style will act as a further obstacle to the user's ability to form desirable social relationships with conversational partners. It appears listeners prefer an SGD to have a voice that is consistent with the characteristics of the person who is using it (Creer et al., 2013). *Pleasantness* focuses on the pleasantness or agreeableness of the synthetic voice (Sefara et al., 2019). Although having a high quality and natural-sounding voice reduces listener fatigue and therefore contributes to positive attitudes towards the SGD by both the user and their conversational partners (Creer, 2009), it is also noteworthy to mention that listeners with greater exposure to synthesised speech, through practice effects, become better at analysing the acoustic–phonetic information in the signal, making better use of the acoustic cues in the synthesised speech, which results in increased intelligibility (Koul & Hester, 2006).

A decade ago, Yamagishi et al. (2012, p. 1) outlined that "it is not easy, and certainly not cost effective, for manufacturers to create personalized synthetic voices". However, technology has sufficiently matured to realistically emulate individual speakers' voices electronically (Watt et al., 2019). Researchers have shown that the results of new speech-synthesis systems are good enough to mislead listeners to thinking that they are listening to authentic voices when they are in fact synthetic (Terblanche et al., 2021; Wang et al., 2020; Wester et al., 2015). Additionally, many speech-synthesis and voice-conversion technologies have become easily accessible through open-source software, and this technology is advancing quickly. Thus, it is not improbable to consider that freely available software be used to assist individuals with CCN. However, these speech-synthesis systems are currently available for various major languages, such as English and other European languages, but are limited for South Africa's indigenous languages (Sefara et al., 2019). Accents also reflect the market size and English speech is therefore often US-accented (Yamagishi et al., 2012). South Africa has 11 official languages, most of which could be considered low-resource languages (de Wet et al., 2017). Thus, South Africa's rich linguistic and social diversity is not well represented in open-source technology.

This is also true for commercially available SGDs, where the speech output is limited. Voices representing a *male*, *female* and in some cases, a *child* are available on the devices, but unfortunately, these same voices are used for numerous individuals and may only represent an individual's identity in a handful of cases (Mills et al., 2014).

Moreover, speech-synthesis technology has customarily focused on adult rather than child speech. Although several methods have been introduced, for a long time, researchers used speaker adaptation in conjunction with hidden-Markov-model (HMM)-based speech synthesis (Cosi et al., 2014; Govender & de Wet, 2016; Hagen et al., 2009; Kumar & Surendra, 2011; Watts et al., 2010). With HMM-based speech synthesis, statistical acoustic models for spectral, excitation, and duration features can be precisely adapted from an average-voice model (derived from other speakers) or a background model (derived from one speaker) using a small amount of speech data from the target speaker (Yamagishi et al., 2010). Novel utterances are then created when models are concatenated (overlapping and adding the signals together), generating the most suitable sequence of feature vectors from the concatenated model for which the speech waveform is synthesised (Yamagishi et al., 2010). Although HMM-based speech synthesis has been used for child-speech synthesis (Govender et al., 2015; Govender & de Wet, 2016; Kumar & Surendra, 2011; Watts et al., 2010), the consensus in the community is that deep neural networks (DNNs) are more suitable for child-speech synthesis (Cosi, 2015; Fainberg et al., 2016; Giuliani & BabaAli, 2015; M. Qian et al., 2016; Y. Qian et al., 2016; Serizel & Giuliani, 2014, 2016; Tong, Chen, et al., 2017; Tong, Wang, et al., 2017). However, there are several difficulties associated with collecting appropriate child-speech data. Govender et al. (2015) believe this is due to (a) children's short attention spans in recording sessions, (b) children's typical articulatory inaccuracies, hesitations and disfluencies, (c) children's limited reading skills as recordings are typically made from read speech, (d) children's fluctuations in emotional expression, and (e) background noise in recording environments. As children need to feel comfortable in recording settings, recording studios are often not appropriate (Govender et al., 2015). Thus, collecting the large amounts of child-speech data necessary for child-speech synthesis is often a challenging process. Despite the challenges, the potential benefits of using this technology for children who have CCN is undeniable. This scoping review summarises the evidence base related to developing synthesised speech for children, and the results are discussed in terms of the implications for service provision for children with CCN.

Methods

Eligibility Criteria

There were several criteria identified for the inclusion of studies in the review. Studies included in the review were those that (a) were published between January 2006 and September 2021 and (b)

included child participants or voices of children aged between 2–16 years old. Selected studies were not limited by design or language, and grey literature was included to ensure a comprehensive review. According to Colquhoun et al. (2014), a scoping review outlines the research to date in a particular field of study, ultimately summarising the research findings and identifying the gaps in the literature. Although the studies may be identified systematically, a scoping review aims to identify all the relevant literature, no matter the study design and/or study quality. We chose to exclude studies older than fifteen years as speech-synthesis technology has substantially changed since then, and the voice output in SGDs often yielded poor results pre-2006, particularly considering speech naturalness and intelligibility. For example, Hoover et al. (1987) reported that genuine speech was more intelligible than synthesised speech in 1987, whereas in 2020, Wang et al. (2020) stated that state-of-the-art text-to-speech systems have the capability to produce synthetic speech that is perceptually indistinguishable from genuine speech by human listeners. Thus, we included current speech-synthesis systems for ecological validity and because older systems were often less intelligible, which made it challenging to generalise and compare the study results to newer systems. Finally, we chose children aged between 2–16 years to reflect the typical school-age population that would likely make use of SGDs.

Search Procedures

The scoping-review protocol (available on request from the corresponding author) was drafted using the preferred reporting items for scoping reviews described by Colquhoun et al. (2014). According to Colquhoun et al. (2014), the steps to conducting a scoping review should follow the Arksey and O'Malley framework stages for the conduct of scoping reviews combined with the Levac et al. (2010) enhancements. The following five steps were included in the process: (1) identify the research question, clarify, and link the aim of the research with the research question; (2) identify relevant studies by balancing feasibility, breadth and comprehensiveness; (3) select studies using an iterative team approach to study selection and data extraction; (4) chart the data using a descriptive analytical method; (5) collate, summarise and report the results (Colquhoun et al., 2014).

A composite search strategy was executed to ensure all studies meeting the selection criteria were identified, whilst avoiding a biased evidence base. The search strategies were drafted by the first author in consultation with a colleague with expertise in electronic-search strategies and further refined through team discussion. The databases and search terms used are outlined in Table 1. Firstly, electronic searches of several databases were conducted using the search terms presented in Table 1. Next, ancestral searches of references cited in studies that met the selection criteria were conducted, which subsequently yielded additional journals for electronic searching. Journal articles, book sections, conference proceedings, conference papers, thesis papers and reports were included in the search.

Irrelevant publications, duplicated publications and publications that did not meet the inclusion criteria were removed. Following this, the final search results were exported into EndNote.

Two reviewers independently screened the 217 publications using Rayyan (Ouzzani et al., 2016) to reduce bias and improve reliability. Both a PhD student in speech and language pathology and a reliability agent, a professional with a doctorate in speech and language pathology, electronically searched the journals and selected relevant studies. Reviewers evaluated the titles, abstracts, and then the full text of all publications identified by our searches for potentially relevant publications. Researchers resolved disagreements on study selection and data extraction by consensus and discussion with other reviewers if needed. Inter-rater reliability was 96.6% for the search procedures (agreements divided by agreements plus disagreements with the outcome multiplied by 100). All disagreements on study selection and data extraction were subsequently resolved through discussion and consensus, and 100% agreement was reached.

Table 1

The search procedures used in the scoping review.

Databases	• EBSCO Host • Scopus • PubMed • Google Scholar
Sources of evidence	• Electronic databases • Reference lists • Electronic searching of key journals • Existing networks, relevant organisations, and conferences
Search terms	(child* AND (“speech synthesis” OR “synthetic voice*” OR “speech synthesi?er” OR “digiti?ed speech”))
	(child* AND VOCA AND (“digiti?ed speech” OR “synthesi?ed speech” OR “speech synthesis”))
	(child* AND “speech generating device” AND (“digiti?ed speech” OR synthesi?ed speech” OR “speech synthesis”))

Coding Procedures

A data-charting form was jointly developed by two reviewers to determine which variables to extract. The data-charting form was piloted with several abstracts and then amended to more accurately capture elements of the speech-synthesis systems. The revised data-charting form was used to chart full-text data. Data charting was independently performed by one reviewer and 15% of the extracted data were audited by a second reviewer, who has their doctorate in speech and language pathology. The results were discussed between reviewers, and discrepancies resolved via consensus. Each study was coded with respect to (a) the aim of the study, (b) the design, (c) the voice-output language, (d) the study population and sample size (i.e., sample size, sex, age, typically developing children/children with disability), (e) the method (novel synthesis system (i.e., hidden-Markov-model-based synthesis, direct waveform concatenation, etc.), commercial SGD (i.e., commercial software and application voices) and review (i.e., scoping review)), and (f) the outcomes.

Results

In total, 58 studies were identified for the review (included studies are asterisked in the reference list). PRISMA guidelines were used. Figure 1 depicts the scoping-review process. Of the 58 included studies, 2 articles were reviews, 5 articles discussed commercially available synthesis systems, and the remaining 51 articles discussed novel synthesis systems. A total of 31 studies focused on English voices, 21 studies discussed languages other than English, and 9 studies did not specify the language used (some studies considered more than one language).

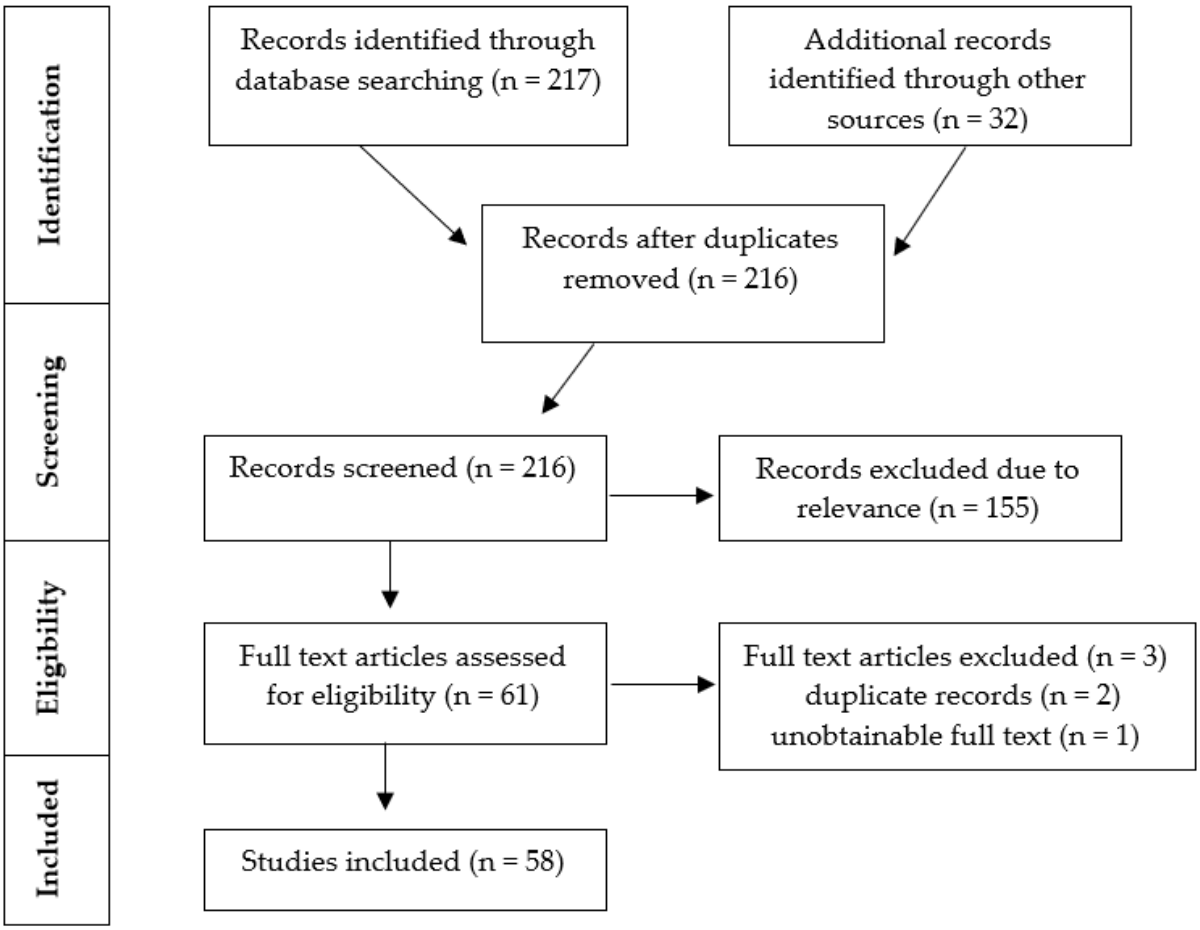


Figure 1

A PRISMA flow diagram depicting the scoping-review process.

Language

Even though the review did not exclude studies by language, it was unsurprising that English was most often used in commercially available voices, such as DECtalk and VeriVox (Von Berg et al., 2009). It was also the language most often selected to create synthesised speech for children. Accents often reflect the market size (Yamagishi et al., 2012), which is likely why the synthesised speech was most often US-accented (Fringi et al., 2015; Govender et al., 2015; Govender & de Wet, 2016; Gray et al., 2014; Hagen et al., 2009; Liao et al., 2015; M. Qian et al., 2016; Shivakumar & Georgiou, 2020). However, it was refreshing to discover studies focusing on other variants of English such as Irish English (Murphy et al., 2020) and Indian English (Kumar & Surendra, 2011; Tulsiani et al., 2017). Although less common, researchers also considered other languages when experimenting with child-speech synthesis, including Norwegian (Begnum et al., 2012), Spanish (Hagen et al., 2009), Punjabi (Hasija et al., 2021), Finnish (Karhila et al., 2012), German (Pucher et al., 2015; Vaz et al., 2009),

Czech and Slovak (Přibilová & Přibil, 2006), Mandarin (Tong, Chen, et al., 2017; Tong, Wang, et al., 2017) and quite often, Italian (Cosi, 2009; Cosi et al., 2014; Cosi, 2015; Gerosa et al., 2007; Gerosa, Giuliani, & Brugnara, 2009; Gerosa, Giuliani, Narayanan, et al., 2009; Giuliani & BabaAli, 2015; Matassoni et al., 2016).

Speech-Synthesis Systems

Prior work has focused on speaker adaptation in conjunction with hidden-Markov-model (HMM)-based speech synthesis (Cosi et al., 2014; Govender & de Wet, 2016; Hagen et al., 2009; Kumar & Surendra, 2011; Watts et al., 2010). In one study, voices created by the HMM-based systems were generally regarded as more similar to the original speaker than the voice-converted unit-selection systems (Watts et al., 2010). Researchers concluded that although the child-speech data had poor coverage of the phonetic/prosodic units of the language, an inconsistent reading style and imperfect recording conditions, it was feasible to build child voices by using the HMM-based speech-synthesis framework (Watts et al., 2010). Similarly, the HMM-prototype voice created in another study had challenges related to intonation and pronunciation, noise levels, naturalness and volume levels (Begnum et al., 2012), but the voice created was regarded to be around seven years old, which suggests the method can build child voices. Jia, Zheng, and Sun (2020) recently attempted to synthesise emotional children's speech by comparing three models (1. HMM model, 2. conditional GAN model, 3. current model CycleGAN). It was clear that the emotional-classification performance of the CycleGAN model was the best, and the accuracy of the HMM-model was the lowest. It appears that although the HMM may be a feasible method for child-speech synthesis, findings suggest that the naturalness of the synthetic speech may be compromised.

Modern synthesis methods, specifically related to deep learning, have had more success at improving the quality of child-speech-synthesis models. Traditionally, researchers have focused on Gaussian mixture models–hidden Markov models (GMM–HMM) for synthetic child-speech development, although more recently, there has been consensus in the community that deep neural networks (DNNs) are suitable for child-speech synthesis (Cosi, 2015; Fainberg et al., 2016; Giuliani & BabaAli, 2015; M. Qian et al., 2016; Y. Qian et al., 2016; Serizel & Giuliani, 2014, 2016; Tong, Chen, et al., 2017; Tong, Wang, et al., 2017), and results show that DNNs have outperformed the older Gaussian mixture models (GMMs) (Metallinou & Cheng, 2014; Shivakumar & Georgiou, 2020). According to Shivakumar and Georgiou (2020), the success of DNNs is due to their ability to use large amounts of training data. Additionally, human speech is filled with non-linearity, and DNNs are able to better approximate the non-linear functions that are needed to model speech (Shivakumar & Georgiou, 2020). Metallinou and Cheng have gone so far as to say that DNN acoustic models work better than GMMs, even when GMMs are trained with approximately eight times more data (Metallinou

& Cheng, 2014). Both Cosi (2015) and Giuliani and BabaAli (2015) used a hybrid DNN–HMM-based automatic-speech-recognition system. They had approximately 10 h of Italian child speech, and results showed improvements over the traditional GMM-based systems.

In addition, using a combination of adult and child training data in the DNN-based models resulted in improved results (Fainberg et al., 2016; M. Qian et al., 2016; Y. Qian et al., 2016). In particular, combining child speech with adult female speech has been proven to be advantageous (M. Qian et al., 2016; Saheer et al., 2013). As the length of an adult female vocal tract is closer to the length of a child's vocal tract in comparison to that of males, manipulating an adult female voice into a child's voice has often been more successful (M. Qian et al., 2016). In fact, the child voice that was adapted from an average-male-voice model experienced significantly larger distortion than that adapted from the average-female-voice model (Saheer et al., 2013). Similarly, a model trained on a male voice resulted in a less naturalistic child voice (Mousa, 2011; Přibilová & Přibil, 2006). Additionally, in order to improve the acoustic mismatch that often occurs as a result of using adult training data on child models, pitch normalisation has been introduced. It has been shown that significant improvement can be obtained with the maximum-likelihood-based explicit pitch normalisation of children's speech (Ghai & Sinha, 2010c, 2010b, 2010a, 2015; Shahnawazuddin et al., 2016; Sinha & Ghai, 2009).

There are several other adaptation techniques that could be utilised to account for the mismatch, and these techniques include, although not exhaustively: adapting acoustic models with maximum-likelihood linear regression (MLLR), constrained structural maximum a posteriori linear regression (CSMAPLR), maximum a posteriori (MAP), speaker-adaptive training (SAT) based on constrained MLLR (CMLLR), and vocal-tract-length normalisation (VTLN) (Shivakumar & Georgiou, 2020). VTLN aims at reducing the interspeaker and inter-age-group acoustic variability due to vocal-tract-length (and shape) variations among speakers by warping the frequency axis of the speech-power spectrum (Gerosa et al., 2007). There are two ways to account for the differences in vocal-tract length between adults and children. One could apply VTLN to adult utterances during training to make the normalised features more like children's speech. Alternatively, one could apply VTLN to the child's utterances during training and testing to make them more similar to the adults' speech (M. Qian et al., 2016). Saheer et al.'s (2013) study shows that combining VTLN with CSMAPLR resulted in an improved adaptation method for both HMM-based automatic speech recognition and text-to-speech. It was clear that when there was a limited amount of adaptation data available, even as little as one sentence, the VTLN yielded the best naturalness and intelligibility results. When only one sentence was used, the CSMAPLR transformation was not intelligible at all. This is an indicator that VTLN is useful in child-speech synthesis. Although VTLN has been recommended for children's speech, it has also been suggested that it should be conducted differently for children when compared to adult speech (Umesh et al., 2007). The results from Umesh, Sinha and Rama Sanand's (2007) study suggested that

unlike conventional VTLN, it was better not to scale the bandwidths of the filters, but rather to scale the filter centre frequencies (Umesh et al., 2007).

Interestingly, aside from creating a child voice by adapting an average adult voice or an average child voice, Karhila et al.'s (2012) study compared two additional adaptation methods using stacked transformations: StA and StVA. In the first method, StA, an average voice trained from adult data was adapted using training data of the average child voice. The resulting synthetic voice sounded child-like with regards to pitch, pronunciation, and rhythm. This model was then further adapted to resemble a specific target child speaker. In the second method, StVA, an average voice trained from adult data was adapted using training data of the average child voice, and then VTLN occurred (Karhila et al., 2012). It was found that stacked transformation systems (StA and StVA) were preferred by listeners and resulted in better adapted voices than directly adapting the average adult voice or the child voice (Karhila et al., 2012).

Child-Speech Data

Due to the scarcity of child-speech data available for researchers, along with the typical articulatory errors, simplifications, hesitations, and disfluencies present in child speech, researchers have attempted to overcome these difficulties when developing child-speech synthesis, in a number of ways. Serizel and Giuliani (2014, 2016) used a similar technique to Tong et al., (2017) to account for limited training data. Hasija et al., (2021) decided to account for the scarcity of available child speech by combining two corpora, one synthesised and one clean (authentic), in order to generate high-quality synthetic child speech. Thereafter, a corpus with a greater quantity was used. Their results indicated that the merged-data corpus showed a reduced word-error rate of the automatic-speech-recognition system with a relative improvement of 9–12%. In a study by Govender, Nouhou, and De Wet (2015), data were selected from an automatic-speech-recognition corpus to build child voices instead of using text-to-speech data. Speech recordings used for automatic-speech-recognition corpora are usually shorter and more spontaneous in comparison to the carefully articulated recorded speech used for text-to-speech development. The criteria considered for adaptation in their study (Govender et al., 2015) included: (1) clean data (with regards to transcription) and no mispronunciations or mistakes in the recordings, (2) data including mispronounced words, (3) number of words in the utterance, (4) rate of speech, and (5) maximum fundamental frequency (Govender et al., 2015). The results showed that when comparing intelligibility, the word-error rates were not as close to voices that were built using speech data that were specifically recorded for speech-synthesis purposes, such as text-to-speech data. However, in terms of naturalness, if data were selected according to particular criteria, then automatic-speech-recognition data could be used to develop child voices that are comparable to text-to-speech voices (Govender et al., 2015).

In addition, many researchers have attempted to build child-speech models by adapting adult-speech models. This has been proven to be a viable method when there are limited speech data available (Begnum et al., 2012; Cosi, 2009; Fainberg et al., 2016; Govender et al., 2015; Hagen et al., 2009; Karhila et al., 2012; M. Qian et al., 2016; Y. Qian et al., 2016; Shivakumar & Georgiou, 2020). In a study by Hagen, Pellom and Hacioglu (2009), a synthetic children's model was derived without child-speech data by using adult-speech data. While they assumed that the availability of child-speech data would have improved the resulting acoustic models, the approach was effective when child-speech data was not available (Hagen et al., 2009). According to a number of researchers (Hasija et al., 2021; Kumar & Surendra, 2011; Shivakumar & Georgiou, 2020), increasing the amount of training or adaptation data results in lower word-error rates. Thus, when creating child voices from adult-speech models, it appears that using more data yields better results, but data should not simply be selected blindly. Some researchers have recommended that one should select training speakers that are closer to the target speaker to train the initial models (Govender et al., 2015).

Shivakumar and Georgiou (2020) have concluded that any amount of child data is helpful for adaptation. In their experiments, even as little as 35 min of child-adaptation data were found to yield relative improvements of up to 9.1% over the adult model (Shivakumar & Georgiou, 2020). However, researchers still need to carefully select the child-speech data, as the typical articulatory errors and disfluencies that are commonly present in child speech will also appear in the synthetic speech without careful selection. This was the case in Kumar and Surendra's (2011) study, and although the output sounded like child-read speech, it was not fluently spoken, which is not ideal for a SGD. Of course, if there are enough child-speech data available, a child voice can be created without directly adapting an average adult voice, as was the case with Karhila et al.'s (2012) study. However, it was found that the adaptation of an average adult voice was preferred to the adaptation of the child voice when there were enough adaptation data. In contrast, when there were very little data available, the child voice was preferred over the average adult voice (Karhila et al., 2012). This appears to support the premise that more data are not always preferred when proper selection does not occur.

When an average voice is created (which is derived from many speakers), one needs to determine how the voices are going to be clustered. In Govender, Nouhou, and De Wet (2015), it was shown that using a gender-independent average-voice model (male and female) resulted in higher quality synthetic child speech than using a gender-dependent average-voice model (male or female). In Watts et al. (2010), two gender-dependent average-voice models were first trained using speaker adaptive training (SAT). Following this, the parameters of both gender-dependent models were clustered and tied using decision-tree-based clustering, where gender was included as the context feature. Lastly, the clustered HMMs were re-estimated using speaker adaptive training, with regression classes for the normalisation being determined from the gender-mixed decision trees (Watts et al., 2010). However, in contrast,

gender-dependent average-voice models were used by M. Qian et al., (2016) and Saheer et al., (2013) where improved performance, as compared to the baseline, was reported.

Intelligibility

It is clear from the data gathered that the speech output of commercially available devices is limited. The speech output of commercially available devices does not often reflect the user's language, age, sex, or personality (Jreige et al., 2009; Koul & Clapsaddle, 2006; Mills et al., 2014; Von Berg et al., 2009). In a study by Begnum et al. (2012), parents of children using SGDs responded positively towards voices that they believed matched their child's identity, particularly as they were sex and age appropriate. However, as the synthesised prototype child voice had flawed intonation and periods of unintelligibility, they reported that they would rather opt for an adult voice that was clear and easy to understand. This was supported by the teachers who needed to understand the voice in demanding surroundings (Begnum et al., 2012). Comparatively, one of the child users stated that he would have liked to use the child voice on his SGD, but that he did not like the sound of the prototype voice created (Begnum et al., 2012). Begnum et al. (2012) found that both the children and their communication partners would prefer a child voice that matches the child's vocal identity but not at the cost of intelligibility. A higher-quality speech output may be more important than matching the child's vocal identity (Begnum et al., 2012).

There are additional factors that may affect the intelligibility of the speech output. According to a study by Drager et al. (2006), two contextual variables (words vs. sentences and topic cues vs. no topic cues) interacted with speech type (digitised vs. synthesised speech). Listeners experienced increased intelligibility for contextual words and sentences and in particular, increased intelligibility of sentences compared to single words. In other words, children who communicate with other children that make use of a single word AAC device may find it difficult to understand as they have little contextual information to point them in the right direction (Drager et al., 2006). This was also found in other studies (Jacob & Mythili, 2008; Von Berg et al., 2009) as listeners reported increased intelligibility of words when words where embedded in sentence utterances rather than in isolation.

Age

Although technology has changed substantially since 2009, results from Von Berg et al.'s (2009) study suggested that child voices in the commercially available systems, DECtalk and VeriVox, were significantly less intelligible than the adult voices in the same commercially available systems, for both single-word and phrase-level intelligibility tasks. Similarly, Shivakumar and Georgiou (2020), conducted an analysis of large-vocabulary continuous-speech-recognition (LVCSR) adaptation and transfer learning for children's speech, using five different speech corpora. The trend showed that as

the age of the child increased, a smaller amount of adaptation data was required. The overall performance increased as the child's age increased, irrespective of the adaptation configuration. Younger children therefore needed more data to reach the same level of performance as older children. In other words, older children had a decreased word-error rate as opposed to younger children (Shivakumar & Georgiou, 2020). In another study focusing on child-speech synthesis, various acoustic adaptation and normalisation techniques were implemented. The word-error rate decreased with age from 6 years to 11 years, resulting in an approximate linear increase in performance with age classes (Shivakumar et al., 2014). Despite this, Drager and Finke (2012) found that a child speaker's original speech intelligibility and articulation skills may be better indicators than age. Some discretion should be applied, as on occasion, a four-year-old child may present with fewer articulatory errors than a seven-year-old child, and therefore appear more intelligible, despite the difference in age.

Discussion

This scoping review addressed the current state of knowledge regarding the development of child-speech synthesis. Based on the reviewed studies, it is clear that child-speech synthesis is still a growing field. However, relative to adult-speech synthesis, developing child speech is notably more challenging for researchers. It is even more challenging when one considers creating synthetic child voices for children with CCN, particularly for those speaking low-resource languages. Thus, these findings are considered in terms of the implications for service provision for this group of individuals.

Language

The review shows that English, particularly US English, is frequently used as the language of choice in child-speech-synthesis systems. However, English is often used due to the market size (Yamagishi et al., 2012) and the availability of English data required for training. Good-quality recordings, in addition to the phonetic and linguistic knowledge that is required for the annotated text resources in the language, can come at a high cost and unfortunately, they do not exist for some languages (Anumanchipalli & Black, 2010). Some researchers have attempted to overcome these difficulties in adult-speech synthesis by crowdsourcing speech samples (Gutkin et al., 2016), using available data such as audiobook data (de Wet et al., 2017) and by using nonideal corpora. In some cases, where a bootstrapping technique can be used, data from a major language can be shared with data from a new language if the new language is comparable with a major language. In those cases, researchers can create synthetic voices for low-resource languages with a small amount of training data (Yang et al., 2015). For example, since Mandarin and Tibetan belong to the Sino-Tibetan language family, the speech data and speech-synthesis framework can be borrowed from Mandarin, whilst only making use of a small amount of Tibetan training data (Yang et al., 2015). This could potentially be a

useful technique for the building of synthetic voices in the Bantu languages of South Africa. In a recent study, results were encouraging when working with data from children speaking in a second language or a non-native language, i.e., Italian children speaking both English and German and German students speaking English (Matassoni et al., 2018). A multi-lingual data adaptation in transfer learning and multi-task learning framework was found to be useful (Matassoni et al., 2018). Nevertheless, there are an overwhelming number of languages that have yet to be considered for speech synthesis.

As English is often prioritised for speech synthesis, children who do not speak English, along with those who do not use it as a first language, are often disadvantaged. This is particularly apparent in South Africa because if a child has to use a pre-loaded voice through an AAC device, English could very well be the child's second or third language (Tönsing & Dada, 2016). As children with CCN may also have comorbid language difficulties, the limited language options subsequently place an additional barrier on their communication. It is well established that an individual's voice is unique and can signify particular elements, such as their physical size, age, sex, race, intellectual ability, geographical and social background, as well as their personality (Jreige et al., 2009; Mills et al., 2014; Sutton et al., 2019). In many countries, including South Africa, a person's language also plays a role in their identity by representing their culture. Thus, the effect that a language selected for an SGD could have on a child's identity should not be minimised. It is also important to remember that for children with CCN who need to make use of SGDs, speech synthesis not only forms the basis of their personal identity, it is also crucial for communication and an essential component of social interaction (Yamagishi et al., 2012). If the language on their device is different from those in their immediate social environment, then the child's communication effectiveness is negatively influenced.

Speech-Synthesis Systems

When considering building child-speech-synthesis systems, one needs to ask several questions: (a) What kind of average-voice model (gender-dependent or gender-independent) is the most appropriate initial model from which adaptation will occur? (b) Are there enough training data for the initial model? (c) Are there enough adaptation data to improve adaptation performance? (d) Should some training/adaptation data be excluded from the model? If so, how much? (e) What kind of adaptation techniques (VTLN, SAT etc.) will be utilised? If so, would it be conducted on the adult speech, the child speech or both? (f) What kinds of transform functions are appropriate?

It appears that there are numerous synthesis methods available, each with their own advantages and disadvantages. As mentioned previously, neural-network-based text-to-speech systems have been gaining popularity in recent years. This speech-generation method is based on deep learning, and it can mine the potential correspondence from multiple corpora and automatically learn the dependence from

the source sequence to the target sequence (Jia et al., 2020). A neural-network-based acoustic model predicts a sequence of acoustic features, including the mel-cepstral coefficients (MCCs), interpolated fundamental frequency and voicing flags. Once completed, a vocoder analyses and converts these MCCs and the fundamental frequency into a waveform, which forms a synthesised voice (Hasija et al., 2021; Wang et al., 2020). According to Wang et al. (2020) and Terblanche, Harrison, and Gully (2021), new neural-network-based text-to-speech systems, such as Tacotron 2, produce synthetic speech that has a perceptually high level of naturalness and good similarity to adult target speakers. Tacotron was also used by Hasija et al. (2021) for the development of children's synthetic speech. DNNs are therefore suitable for child-speech synthesis (Cosi, 2015; Fainberg et al., 2016; Giuliani & BabaAli, 2015; M. Qian et al., 2016; Y. Qian et al., 2016; Serizel & Giuliani, 2014, 2016; Tong, Chen, et al., 2017; Tong, Wang, et al., 2017) and often outperform the older GMM models (Metallinou & Cheng, 2014; Shivakumar & Georgiou, 2020). However, it should be noted that a large amount of training data is not always freely available when developing child speech, particularly when one considers building child speech for low-resource languages. It appears that in order to account for this, researchers began using both adult and child speech for training in the DNN-based models, and this resulted in improved results.

There are a number of techniques that have been introduced to improve child-speech synthesis. Firstly, researchers need to define better acoustic features for children's speech. The most commonly used features include mel-frequency cepstral coefficients (MFCCs), filterbank, and perceptual-linear-prediction (PLP) coefficients (M. Qian et al., 2016). Typically, MFCCs achieve the best performance in GMM-based systems (Ghai & Sinha, 2015; Shivakumar et al., 2014) while mel-filterbank coefficients are often used in DNN-based systems (M. Qian et al., 2016). Secondly, due to the differences in child-speech development, pronunciation modelling is required (Shivakumar et al., 2014). Thirdly, VTLN is often used to account for the differences in vocal-tract length between speakers (Gerosa, Giuliani, & Brugnara, 2009; Ghai & Sinha, 2009; Karhila et al., 2012; Saheer et al., 2013). Fourthly, modal-adaptation techniques are often used, such as maximum a posterior (MAP) and maximum-likelihood linear regression (MLLR) (Gerosa et al., 2007; M. Qian et al., 2016).

Child-Speech Data

As previously discussed, collecting child speech presents with many difficulties. The type of child speech typically available in corpora does not always provide complete coverage of all the speech units in the language, and it is often inconsistently read and imperfectly recorded (Govender et al., 2015; Kumar & Surendra, 2011). According to a number of researchers (Hasija et al., 2021; Kumar & Surendra, 2011; Shivakumar & Georgiou, 2020), increasing the amount of training or adaptation data results in lower word-error rates, indicating greater intelligibility, even if imperfect data are included (Govender et al., 2015). This is in contrast to another study, where researchers concluded that using

fewer data of superior quality is preferred in adult-speech-synthesis models, as opposed to using more data of inferior quality (de Wet et al., 2017). Although fewer data are available, the high-quality data reportedly result in adult voices of improved naturalness and intelligibility (de Wet et al., 2017).

Comparably, when adapting adult-voice models to resemble child speakers, some researchers suggest clustering the data (i.e., age, sex, max fundamental frequency, etc.) so as to develop an average-voice model that closely matches the target child speaker to some degree (Govender & de Wet, 2016; Jreige et al., 2009), while other researchers suggest adaptation can be performed after the average-voice model has been trained, with adaptation techniques such as VTLN (Gerosa, Giuliani, & Brugnara, 2009; Karhila et al., 2012; Saheer et al., 2013). In a study by Yamagishi et al. (2010), it was seen that the greater the difference between the average-voice model and the target speaker, the poorer the resulting target voice quality would be. This finding suggests that simply using an average adult voice for adaptation to a child target speaker will yield poor results. However, it is quite clear from the review that this method has been used on numerous occasions (Begnum et al., 2012; Fainberg et al., 2016; Govender et al., 2015; Hagen et al., 2009; Karhila et al., 2012; M. Qian et al., 2016; Y. Qian et al., 2016; Shivakumar & Georgiou, 2020). In order to pre-empt the decreased quality due to the mismatch between adult and child, researchers have suggested selecting training speakers that are closest to the child target speaker. The resulting quality could be improved by training a large amount of pre-selected data, aided by a neural-network classifier, to better match the children's speech (Liao et al., 2015). Training speakers to resemble the child's vocal quality to some degree could either be done in the initial training phase by selecting similar training speakers, or after data normalisation has occurred. In addition, if the average adult voice is further augmented with a small amount of children's speech, a closer match may be found (Hasija et al., 2021; Shivakumar & Georgiou, 2020).

Intelligibility

Because of the lack of redundant auditory and visual cues found in synthesised speech, listeners must allocate more attentional resources to process synthesised speech, as compared to natural speech (Drager, Reichle, et al., 2010). These factors, such as a lack of visual cues, are likely to influence the comprehensibility of any communication mode (natural or synthesised), when the speaker cannot be seen, such as what one might experience with telephone speech (Drager et al., 2004). Similarly, natural speech is often comprehended faster than synthesised speech, especially when there is a high level of background noise or the listener's attention is divided (Drager, Reichle, et al., 2010). However, as the quality of synthetic speech improves, the margin between synthetic speech and natural speech decreases, as does the cognitive load required (Creer, 2009). In terms of what individuals prefer, it appears that people prefer listening to voices that match the vocal identity of the user (Begnum et al., 2012; Creer, 2009; Gorenflo et al., 1994). Despite this, there is evidence to suggest that individuals

would rather choose high intelligibility of the speech output instead (Begnum et al., 2012; Drager et al., 2004; Gorenflo et al., 1994). A high level of intelligibility is therefore crucial in the development of synthetic speech.

It has been shown that human listeners adjust in subtle but systematic ways to understand synthetic speech. Understanding synthetic speech often requires a greater cognitive load than understanding natural speech, but this cognitive load decreases when the listener becomes accustomed with the voice (Creer, 2009; Drager, Reichle, et al., 2010). For children and adults, comprehensibility of the synthesised speech signal often improves after greater exposure (Pucher et al., 2015). This is also true for individuals with intellectual disabilities, as studies showed that their perception of the synthetic speech improved after systematic exposure to it (Koul & Clapsaddle, 2006; Koul & Hester, 2006).

Results have suggested that context and the length of the utterance play a role in the intelligibility of synthesised speech (Drager et al., 2006; Drager, Reichle, et al., 2010). In other words, longer utterances are more intelligible than single words, unless the listeners are given a closed response set (i.e., having a set of predefined answers and pointing to a picture in response to a stimulus) (Koul & Clapsaddle, 2006; Koul & Hester, 2006). Longer utterances typically contain more linguistic context than single words, which listeners can use to increase signal-independent information. This is clinically important as some children with CCN may have comorbid difficulties, such as decreased working memory and impaired language skills, resulting in decreased use and comprehension of longer utterances (Drager, Light, et al., 2010). Thus, single-word AAC devices are often common starting points for children with CCN. However, if their speech output is supported with a symbol or picture on their AAC device, it may assist their communication partners in understanding them. It is well-recognised that successful AAC communication depends on both the AAC user and their communication partners (Creer et al., 2013; Jette et al., 2017; Kent-Walsh et al., 2015; Moorcroft et al., 2019).

Intelligibility of the speech output on a speech-generating AAC device is essential for several reasons. If children are unable to communicate with their natural voice, the speech output will assist them in learning how to use the AAC system; it provides feedback to the child and their communication partner and it allows for successful interaction opportunities with new communication partners (Drager et al., 2006). Moreover, the speech output may provide additional verbal modelling, which could result in an increase in the child's spontaneous speech production (Wendt et al., 2019). Researchers have discussed how sound produced by an iPad or a SGD may act as a reinforcer for a child user, which subsequently motivates them to use it to communicate (Ganz et al., 2012; Schlosser et al., 2007; Wendt et al., 2019). In one study, children with intellectual disabilities were also able to generalise their knowledge of the acoustic-phonetic properties of synthetic speech to novel stimuli (Koul & Clapsaddle,

2006). This clearly has some vitally important clinical implications for children with intellectual disabilities and would likely result in an increase in their participation in academic environments.

Age

Typically, adults are considered more intelligible than children due to the expected patterns of speech simplifications and articulatory errors often observed in a developing child's speech. Along with the physical, linguistic, and prosodic differences observed in children's speech (Hasija et al., 2021; Shivakumar & Georgiou, 2020), it is not surprising that the acoustic and intelligibility mismatch between adult and child speech is also considered a challenge in the development of synthetic speech for children. It is well established that characteristics of speech, such as pitch, formant frequencies and phone duration are related to the age of the speaker (Gerosa, Giuliani, & Brugnara, 2009). These acoustic differences result from children having shorter vocal tracts and smaller vocal folds than adults (Cui & Alwan, 2006). Younger children show higher pitch and fundamental frequency in comparison to older children (Giuliani et al., 2006). In Shivakumar and Georgiou's study (Shivakumar & Georgiou, 2020), the trend showed that as the age of the child speaker increased, less data adaptation was required. Older children did not experience as great of a mismatch between the adult speech (which was used as training data), while younger children showed that considerably more data were needed to account for the mismatch (Shivakumar & Georgiou, 2020). Younger children therefore show a considerably higher amount of intra- and inter-speaker variability as compared to older children and adults (Gerosa et al., 2007; Gerosa, Giuliani, Narayanan, et al., 2009). Younger children have high levels of acoustic complexity, which can be attributed to "three main factors (i) shifted overall spectral content and formant frequencies for children, (ii) high within-subject variability in the spectral content that affects formant locations and (iii) high inter-speaker variability observed across age groups, due to developmental changes, especially vocal tract" (Shivakumar & Georgiou, 2020, p. 2). Thus, more parameters were required to accurately capture the increased complexity in their speech (Shivakumar & Georgiou, 2020). If one were to consider building a young child voice, with its high level of acoustic complexity, it may be prudent to first consider the availability of the data necessary for adaptation. Although there is not a consistent cut-off age for the most suitable levels of intelligibility (Drager, Reichle, et al., 2010) (i.e., seven-year-old speech vs. four-year-old speech), it appears that as a child's level of intelligibility increases and their acoustic variability decreases, which may be at approximately nine years old, the less challenging it would be to build a synthetic voice for them.

Conclusion

This scoping review addressed the current state of knowledge regarding the development of child-speech synthesis based on research conducted over the last 15 years and reveals potential directions for

future research. However, the possibility of not including all relevant articles must be recognised. Selecting other databases may have identified additional relevant studies. Additionally, relevant articles may have used terms other than speech-generating device or voice output communication aid. Finally, we did not critically appraise individual sources of evidence, as the focus of this review was to identify available evidence rather than to evaluate it.

This scoping review did not evaluate specific speech-synthesis systems or judge the utility of these systems for children with CCN. However, what emerges from the evidence is that speech-synthesis technology has improved remarkably over the last 15 years. In fact, in the last few years, it has become possible to create intelligible and natural-sounding synthetic speech that has the potential to mislead listeners to thinking that they are listening to authentic speech. Many speech-synthesis and voice-conversion technologies have become easily accessible through open-source software. However, based on the studies reviewed, relative to adult-speech synthesis, developing child-speech synthesis, particularly for young children, is notably more challenging for researchers. Child speech often presents with acoustic variability, disfluencies, and articulatory errors. In addition to this, it is often challenging to collect child speech due to children's short attention spans, limited reading skills and the diverse recording environments.

To account for these challenges, numerous researchers have attempted to adapt adult-speech models, using a variety of different adaptation techniques. In most cases, adult-speech data are used in combination with a small amount of child-speech data to create a synthetic child-like voice. Adapting adult speech has proven successful in child-speech synthesis and it appears that the resulting quality can be improved by training a large amount of pre-selected speech data, aided by a neural-network classifier, to better match the children's speech.

For children who are unable to communicate using their natural speech, speech synthesis could provide a viable means of communication. The selected synthetic voice used in the speech-generating device is therefore likely to become an extension of themselves. With that in mind, we propose that the synthetic voice be individualised to best represent the child's vocal identity, with regards to at least their language, sex, and age. With an individualised synthetic voice, children with complex communication needs could potentially increase their intentional communication skills, participation, language, and literacy skills in a classroom setting. As multiple children in one classroom may need a synthetic voice, having an individualised voice would benefit them greatly as teachers would be able to differentiate speakers in class, there may be greater technology-adoption rates and an increased level of socialisation between children. Future research could focus on developing a system that is acceptable to the child and improves its performance over time based on continued use by the child.

As language is a large part of an individual's vocal identity, the language selected for the device is another important element to consider. Many of the speech-synthesis systems are usually designed for major languages, such as English, but are limited for low-resource languages. Promisingly, it was discovered that if there are enough training data available, either collected in the typical fashion or atypically, through crowdsourcing, or by combining language-similar corpora (borrowing speech data from other corpora that fall within the same language family), one should be able to create natural and intelligible synthetic speech for children in any language. We therefore believe that future research should investigate individualised synthetic speech for children with complex communication needs, with special attention to children who make use of low-resource languages. It would be interesting to determine if the residual speech produced by a target child with complex communication needs could be combined with speech data from typically developing children and utilised to develop a unique and individualised synthetic voice for a particular target child, whilst making use of open-source speech-synthesis software.

MANUSCRIPT TWO

Challenges, perceptions, and implications of AAC use in South African classrooms: An exploratory focus group study.

Camryn Terblanche, Michelle Pascoe, and Michal Harty

Division of Communication Sciences and Disorders, University of Cape Town, SA

trbcam001@myuct.ac.za/ michelle.pascoe@uct.ac.za /michal.harty@uct.ac.za

Abstract

Communication partners are instrumental in the successful use and implementation of augmentative and alternative communication systems (AAC), especially in schools, but stakeholder views from low- and middle-income countries (LMICs) are not well represented in the literature. Focus group interviews with 7 professionals and 3 caregivers from South Africa were conducted to understand their perceptions and experiences of AAC use and implementation. The results highlighted additional issues which practitioners in LMICs need to consider when implementing AAC in under-resourced schools for learners with special education needs. Although some challenges overlap with those experienced in high-income countries, such as support and training, high staff turnover and burnout, large caseloads, and language and codeswitching differences, these challenges present differently in low-income contexts, requiring alternative solutions. High-income countries do not often need to consider the risk a high-tech AAC device places on the child and their family due to the risk of crime in low-income contexts, the device affordability, the device features, and the range of accents available when introducing an AAC system. This exploratory study suggests that LMICs, like South Africa, could make great strides towards providing appropriate AAC technology for all if i.) strategic partnerships between governmental and non-governmental groups were put in place, ii.) appropriate communication, training and support systems were established, and iii.) evidence-based core-language AAC systems were created.

Index Terms: augmentative and alternative communication (AAC), barriers and facilitators, children, communication partners, complex communication needs (CCN), low-and middle-income countries (LMIC), learners with special education needs (LSEN), South Africa, stakeholder perspectives

Introduction

Effective augmentative and alternative communication (AAC) depends on both the AAC user and their communication partners (Jette et al., 2017; Kent-Walsh et al., 2015; Moorcroft et al., 2019b). AAC is an important option for children with complex communication needs (CCN), reducing barriers to communication, increasing opportunities for language and literacy development, and ultimately improving their quality of life (Leonet et al., 2022). Children with CCN may not develop typical speech, language, and communication skills due to motor, language, cognitive and/or sensory perceptual impairments. AAC is divided into three categories: no-tech, low-tech and high-tech. No-tech AAC uses facial expressions and motor movements, such as manual sign. Low-tech AAC involves basic communication devices, such as paper-based communication boards and books, whereas high-tech AAC includes speech generating devices and other computer and tablet-based technologies to send a verbal message (Leonet et al., 2022). Equipping schools with the tools they need to provide inclusive education, not only assists children with CCN, but also helps children at risk for difficulties with learning and literacy (Kathard et al., 2011). When schools train teachers on the importance and relevance of AAC, whilst providing them with resources to adequately support the children, the children thrive (Moorcroft et al., 2019a). Moreover, peer scaffolding, when peers demonstrate a skill and offer support, assists the child using AAC and their peers. Peers develop language, social and academic leadership skills, whilst also increasing awareness and acceptance of AAC use (Finke et al., 2009).

Traditionally, AAC support has focused solely on individuals with CCN whilst neglecting the needs of caregivers, service providers, peers, and other communication partners (Light et al., 2019). Early AAC intervention is paramount for early communicative success, and intervention effectiveness ultimately relies on acceptance and modelling of AAC by communication partners (Moorcroft et al., 2019b). Despite the need for AAC, challenges often occur which result in the rejection or abandonment of AAC systems by children with CCN and their families. To understand why AAC systems are underutilized, Baxter et al., (2012) conducted a systematic review of barriers and facilitators to the provision and use of high-tech AAC systems between 2000-2010. Moorcroft et al. (2019a) undertook a similar review, considering the barriers and facilitators to the use of low-tech and unaided AAC between 2000-2016. These reviews found that rejection rates dramatically increase if communication partners are not familiar, or do not fully accept the AAC device. The perceptions of speech and language therapists (SLTs), teachers, caregivers and peers about factors that contribute to the acceptance, rejection or abandonment of AAC systems – and their attitudes towards users of such systems – has been discussed (Baxter et al., 2012; Dada et al., 2016; Moorcroft et al., 2019a, 2019b; Tönsing and Dada, 2016; van Niekerk et al., 2019). However, less is known about the specific challenges encountered by stakeholders when introducing AAC systems within special education settings in low- and middle-income countries (LMICs), and potential ways to overcome these challenges.

Despite the evidence indicating that communication partners play a crucial role in the successful use and implementation of AAC systems (Jette et al., 2017; Kent-Walsh et al., 2015; Moorcroft et al., 2019b), there are limited studies considering the perspectives of stakeholders from LMICs (Mukhopadhyay and Nwaogu, 2009; Tönsing et al., 2019; Tönsing and Dada, 2016; van Niekerk et al., 2019; Wormnaes and Malek, 2004). Research from high-income countries often serves as the gold standard for AAC intervention. However, additional challenges are experienced when implementing AAC in under-resourced and low-income contexts (van Niekerk et al., 2019). For instance, there is a lack of AAC knowledge and skills amongst teachers and SLTs in many African countries, such as Botswana, Egypt and South Africa (Mukhopadhyay and Nwaogu, 2009; Tönsing et al., 2018; Wormnaes and Malek, 2004). Less than 50% of the Egyptian participants in Wormnaes and Malek's (2004) study felt that they were sufficiently qualified in the field of AAC, despite working with individuals with little or no functional speech, and the majority of the participants in Mukhopadhyay and Nwaogu's (2009) study in Botswana displayed negative attitudes towards AAC. Tönsing et al., (2018) discussed how the skills of AAC service providers in South Africa are often limited when designing and implementing AAC systems and techniques in multiple languages. In order to provide some background to this study, the South African context will be further described.

The South African Context

South Africa's official language policy allows schools to select any of the 12 official languages as the language for learning and teaching. However, multilingualism, multiculturalism and the diverse range of social issues can make it difficult to ensure equality in all educational spheres (Pascoe and Norman, 2011). Speakers of indigenous languages are often under-served due to the lack of contextually relevant resources in African languages, including AAC, and the language mismatch between professionals and the children/families they serve (Kathard et al., 2011; Pascoe and Norman, 2011). It is estimated that approximately 63% of children in South Africa come from poor households (Statistics SA, 2020), many of whom face a higher level of crime, overcrowded living situations, generational illiteracy, and increased trauma and violence (van Niekerk et al., 2019). These environmental barriers have a negative impact on children's access to education and many of South African children from poor households do not obtain the exit requirements from secondary school which would allow them to make use of tertiary educational opportunities.

In addition, the South African Department of Health (responsible for the implementation of early intervention) is only able to meet 25%-65% of the total assistive products required, despite providing healthcare services to over 80% of the country (Visagie et al., 2020). Once children requiring AAC reach school-going age, they are referred to the Department of Education for continued intervention (van Niekerk et al., 2019). Unfortunately, many children are lost in the system due to

limited resourcing and poor coordination at these transition points (van Niekerk et al., 2019). In addition, up to 70% of school-going age children with disabilities do not even attend school in South Africa. Of those who do attend, most are still in separate special schools for learners with disabilities (Donohue and Bornman, 2014) where both rehabilitation personnel, and suitable assistive technology are not always available (van Niekerk et al., 2019). Although South Africa is moving towards an inclusive education system (South African Department of Education, 2001), the implementation has been slow, largely due to poor policy directives, the large number of students in mainstream classrooms, lack of funding and a lack of support for schools and teachers (Donohue and Bornman, 2014).

With this in mind, strategies used in high-income countries, to provide learners with appropriate AAC technology, may not be applicable in LMICs. Considering stakeholders' roles in the acceptance of AAC, the aim of this exploratory study was to document multiple stakeholders' experiences, perceptions of implementing AAC in their context, and their ideas for overcoming the challenges surrounding the use of AAC within schools for Learners with Special Education Needs (LSEN) in South Africa. The research question was therefore: What are the perspectives of teachers, SLTs, and caregivers on the implementation of AAC systems for children with CCN in South African education settings, and how do these perspectives reflect the challenges and opportunities of AAC implementation in LMICs, such as South Africa? There were three research objectives: 1) To examine the perspectives of teachers, SLTs, and caregivers regarding the current accessibility and effectiveness of AAC systems when faced with varying levels of resources, 2) To explore the challenges faced by these professionals and caregivers in implementing AAC systems in a LMIC, such as South Africa, and 3) To identify the support needed for professionals and caregivers in LMICs, such as South Africa, to improve AAC implementation.

Method

Qualitative approach

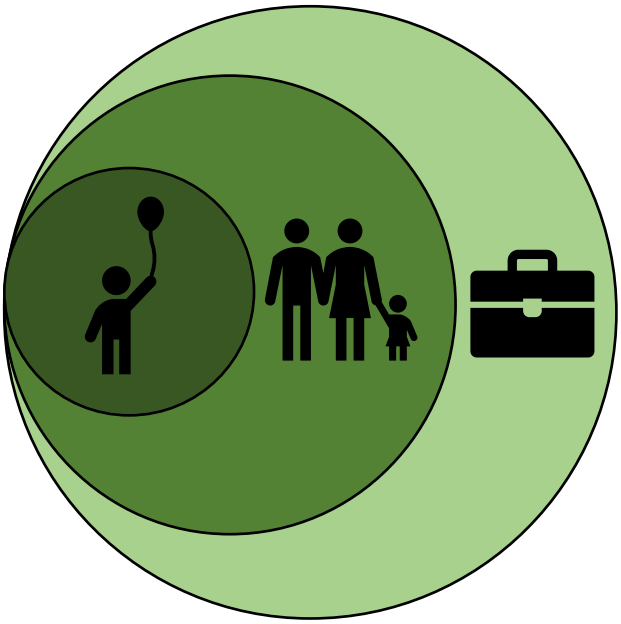
This project is part of a larger study where naturalistic synthetic child speech was created for three children with CCN, in three languages, namely South African English, Afrikaans and isiXhosa (Terblanche et al., 2024). Prior to the participants hearing the synthetic voices, face-to-face focus groups were conducted to determine AAC implementation challenges currently faced by participants. Thus, a descriptive qualitative design was used during this exploratory project as it allowed the researchers to obtain detailed descriptions from several stakeholders (Terre blanche et al., 2006). We anticipated that participants' perspectives would highlight key challenges, resource limitations, and varying views on the accessibility and effectiveness of AAC systems in the context of a LMIC, such as South Africa. This exploratory focus group study does not seek to confirm predefined predictions or reach consensus

on specific points, but rather to gather rich insights into the issues at hand, which will extend our understanding of the topic (Liamputtong, 2011). For this study we used purposive sampling as the participants were selected based on their relationship to the three children for whom the synthetic child speech was created. Participant selection criteria included familiarity with at least one of the 3 child participants, including their specific needs and difficulties; and the ability to participate in focus groups conducted in English. In this study, two focus groups were conducted after working hours, in a school setting. One focus group included caregivers of the 3 children with CCN whilst the other included teachers and SLTs who work with children with CCN. Two focus groups were conducted to ensure that power imbalances did not occur between participants.

Ethical approval for this study was obtained from the Human Research Ethics committee (HREC no. 765/2021) at a South African university and followed the guidelines outlined in the Helsinki Declaration of 2013 (World Medical Association, 2013). The appropriate provincial education department and the relevant school principals gave permission to access the schools. All participants were asked to maintain confidentiality, and each participant had to be legally competent and informed before they were invited voluntarily to sign a consent form.

Participants

Purposive sampling was used, with researchers primarily recruiting participants within 3 LSEN schools in Cape Town. Figure 1 shows how participants were selected. Once the three children were selected for the larger study, the adult participants (professionals and caregivers), were selected by moving outwards in concentric circles. Therefore, the participants were familiar with at least one of the child participants, including their specific needs and difficulties. Participants with different linguistic backgrounds were selected to reflect the multilingual context. Most participants were recruited from three LSEN schools, with the exception two professionals who were recruited via social media.



	Three children with CCN.
	Caregivers/family members of the three children with CCN.
	Teachers and SLTs working with the three children with CCN.

Figure 1

Participant selection process

Caregivers

Six caregivers gave written consent to participate, but three were unable to attend on the day. Thus, only three adult caregivers participated in the focus group ($\bar{x}$ age= 47 years old). Two participants were mothers of children with CCN whilst another was a grandmother of a child with CCN. Although the children were encouraged to use AAC at school, the caregivers were not yet exposed to AAC, as the children were not allowed to take their school AAC devices home. The children's diagnoses included autism spectrum disorder, significant intellectual disability, and cerebral palsy ($\bar{x}$ child age= 10 years old).

Professionals

Twelve teachers and SLTs gave written consent, but only seven were available to participate ($\bar{x}$ age= 37 years old). Participant A consults with LSEN schools, trains SLTs in the area, and has extensive experience with both high- and low-tech AAC systems. Similarly, Participant B previously worked at a company specialising in assistive technology, where she trained caregivers and implemented various AAC systems, and currently works in a LSEN school, where high- and low-tech AAC are available. Participant C has access to both high- and low-tech AAC at her LSEN school but reports that she mainly incorporates low-tech AAC into her sessions. Participant D has experience implementing low-tech AAC in both private practice and her previous LSEN experience. Participant E

mainly incorporates low-tech AAC into her sessions as there is currently only one child in her LSEN school who has access to a high-tech device. Participant F has had high-tech AAC in her LSEN classroom for the last 3 years, following a donation of high-tech AAC systems to her school. Lastly, Participant G has limited experience implementing low-tech or high-tech AAC, although it is available in the LSEN school, as they rely primarily on manual sign. Table 1 outlines the participants' key sociodemographic characteristics.

Table 1

Professional focus group participant demographics (n=7)

Demographics	Number of Professionals
Age range	
20-30	2
31-40	3
41-50	2
Gender	
Male	1
Female	6
Languages used in intervention/classroom	
English	2
Afrikaans/English	2
Afrikaans/IsiXhosa/English	3
Years of experience	
3-5	1
6-10	2
11+	4
Service sector employment	
Public	5
Private	2
Employment type	
Teacher	2
SLT	5

Data collection

The data were collected during two focus group sessions, lasting approximately 1.5 hours each. Separate interview schedules were devised. The facilitator asked topic-related descriptive questions to guide the discussion, starting with general questions (e.g., "What is it like communicating with your clients/children?") and proceeding to the research questions (e.g., "What challenges have you experienced, implementing/using AAC in South Africa?") (Terre blanche et al., 2006). To ensure credible data, focus groups were audio recorded and immediately following each group, typed transcripts were prepared. To maintain anonymity, no identifying information was recorded in the transcripts. Transcripts were provided to the participants for member checking, and none indicated they required changes.

Data analysis

Verbatim transcripts of the discussions were analysed via reflexive thematic analysis in a five-stage framework approach, using NVivo (QSR International, 1998). According to Braun and Clarke (2006, 2019, 2021), reflexive thematic analysis is a theoretically flexible interpretive method that identifies and analyses patterns of meaning in a dataset. In stage one, data familiarisation, the transcripts were reviewed, and patterns were identified. Initial codes were generated in stage two. Codes identify interesting features of the data (semantic content or latent) and refer to the most basic element of the raw data that can be assessed in a meaningful way. In stage three, themes were identified and all relevant extracts from the transcripts were collated and categorised. The themes were refined and reviewed in stage four, e.g., some themes were merged. In stage five, defining and naming the themes, the themes were divided into sub-themes, and the data were analysed and further refined. Two judges independently applied the coding framework to the interview data. Discrepancies were discussed until consensus was reached and 100% agreement between the two judges occurred, ensuring validity and reliability of the analysis and the identification of key information.

Results

Caregiver focus group

Themes and subthemes that emerged after analysing the caregiver focus group are presented in Table 2. Two themes emerged from the data, namely, device suitability and benefits.

Table 2

Identified themes, subthemes, and examples discussed by caregivers.

Themes	Subthemes	Instances identified	Examples provided by caregivers
Device suitability	Education ±	5	Most caregivers want their children to only use verbal language to communicate.
	Language ±	2	The language used at school is not necessarily the same language used at home.
	Safety *	4	High-tech AAC devices place the children and their family's safety at risk.
Benefits	Improvements	5	Verbal language use, including sentence length, has improved since incorporating AAC.
	Attitudes and expectations	6	Peers might not make fun of the children if they think the high-tech device is interesting.

± *Challenges shared with high-income countries but present differently in low-and middle-income contexts.*

* *Challenges unique to low- and -middle income countries.*

Device suitability

The device suitability theme encompasses the appropriateness of using high-tech devices for particular families, and situations. Three subthemes were identified, including education, safety, and language.

Caregivers reported feeling ill-informed about alternative forms of communication. During discussions surrounding AAC knowledge, one caregiver stated, "we have a lot to learn". In addition, some caregivers acknowledge that they have never learnt how to communicate with their children who have CCN: "He cannot speak at all, so he uses [manual] sign… I don't know [manual] sign at all." Another caregiver highlighted a preference for spoken language, "because we [family] don't actually want him to use [AAC]. We want him to speak". This was echoed by another caregiver who said "we also prefer talking to [manual] sign… Otherwise, when are they going to learn to talk?" Caregivers also

indicated that AAC devices should only be used at school, where children are safe and supported, "he will sit with that [AAC device] at school because [the SLT] is there to help him…"

With this in mind, if caregivers don't feel comfortable using AAC at home, then the child's AAC exposure and interaction is limited to the classroom. However, the language used at school is not necessarily the language used at home, which further hinders AAC use at home as one caregiver pointed out, "his teacher mentioned that he speaks English with her [but he] speaks Afrikaans at home". Although many children become multilingual over time due to the variety of languages spoken at home and at school, "they want to [teach] him isiXhosa at school", and "he is Afrikaans…[but] when he is cross, he complains in English", AAC devices are often not representative of this diversity.

Additionally, caregivers raised concerns for their child's safety while using high-tech AAC devices and would prefer the devices stay at school. They are concerned that theft of the device may occur from their home or while on public transport, placing the child and family at risk. One caregiver stated "because outside, they will break the [AAC device]. They will either take it from him or break it…and he can't come back and say that [person] took it or what happened to the [device]". Caregivers were also concerned about the safety of the device software. Typically, passwords are required before updating AAC software, but they believe their children will bypass these safety precautions and delete important components by mistake, "We will be worried [about the AAC software] because [child] is always deleting games and stuff [on our phones]". Finally, due to the broader community's lack of understanding about AAC use, caregivers expressed concern that their children may be bullied, "I think they will make fun of him if he [goes] outside with the [AAC device]." One caregiver mentioned that due to this "he doesn't speak or play with [neighbourhood children] anymore". The caregiver explained that the child only plays "at school with his new friends" where peers are more exposed to different forms of communication.

Benefits

This theme encompasses the benefits experienced by caregivers who have children that use AAC. It has two subthemes, including improvements, as well as attitudes and expectations.

Although some caregivers were hesitant about using the device, one caregiver stated that they noticed improvements in their child's verbal vocabulary and sentence length after using the device at school, "he is catching up nicely now," while another mentioned that "he couldn't speak at first, but he is starting [after AAC]".

One caregiver shared her concerns about the expected poor attitudes of children as "kids might make fun of him" for using the device, but another suggested that having a high-tech device would

likely encourage “other children to play on it or do the same thing” and this would be “exciting”. They expressed that children’s attitudes might change over time as the AAC technology “would be interesting” for them.

Professional focus group

Themes and subthemes that emerged after analysing the professional focus group are presented in Table 3. Five themes emerged from the data, namely, support and training, device and software, education system, language and code-switching, and AAC benefits. The lack of support and training was the greatest challenge described by professionals.

Table 3

Identified themes, subthemes, and examples discussed by professionals.

Themes	Subthemes	Instances identified	Examples provided by professionals
Support and training	Support ±	8	School-based therapists do not feel supported by district personnel, especially regarding AAC.
	Caregiver education ±	10	Perceptions towards AAC and reduced understanding of diagnosis, limit AAC use at home.
	Professional training	17	Professional’s need to be trained on how to implement AAC in the classroom.
Device and software	Safety *	2	High-tech devices are often not sent home because of safety concerns.
	Device affordability *	3	Socioeconomic factors are often a barrier to the use of high technology AAC.
	Device features *	11	Most AAC apps are only available on Apple, rather than Android.

Education system	High staff turnover and burnout ±	3	High staff turnover due to burnout often limits the child's progress.
	Therapist-teacher collaboration	10	Some teachers like the push-in method, whilst others prefer the pull-out method.
	Caseload ±	8	Therapists do not have enough time to create and maintain AAC systems.
Language and code-switching	Language ±	7	There are very few AAC resources in all the South African languages.
	Accents *	5	Synthetic speech (particularly accents) does not reflect the South African population and age group.
	Code-switching ±	2	Current devices are not able to incorporate code-switching.
Benefits	Attitudes and expectations	8	People's expectations change, for the better, once functional AAC use is introduced.
	Identity and personality	4	Children no longer need to feel isolated as they can share their ideas, thoughts, and opinions.

± *Challenges shared with high-income countries but present differently in low-and middle-income contexts.*

* *Challenges unique to low- and -middle income countries.*

Support and training

This theme outlines the imperfect support structures and limited training opportunities often experienced by professionals. Three subthemes were identified, including support, caregiver education, and professional training.

SLTs reported feeling unsupported by the district level support staff within the education system. One therapist shared her experience as a senior SLT in a LSEN school:

> The school didn't employ me; the Department of Education employed me. I had a district therapist, who did not support me in the way that I needed, who did not allow me to support others in the way that I wanted. When I started, all the schools met [at an AAC forum]…. And as soon as funding became involved, the district therapist said, 'you're out of line, this is now district level, we're taking over' and it stopped. The school [was] lovely, amazing, incredible; beyond the walls of the school? Lack of support, like you can't believe.

This lack of support is sometimes also felt at a school level as many have "never had a senior therapist" who could assist with AAC decision making. The few SLTs who could speak to a senior SLT indicated that it felt like "a weight off [their] shoulders". One SLT expressed: "We did our best with what we had but with very little support or anything that came with the AAC. There were probably six kids that would have definitely benefited [from AAC] so it was a shame that there was nothing that we could do." Another SLT described the impact of this lack of support on the sustainability of AAC use:

> Until everyone's not working in individual schools and individual silos trying to make something happen… And the people who employ us, the people who support us, don't start putting systems/support/training in place - It's always going to be individual teachers, individual therapists, trying their level best for years until the next person comes and tries their level best for years.

Educating caregivers about their child's diagnosis and AAC system was another important subtheme. Professionals feel that it may be unrealistic to expect them to start incorporating AAC immediately after receiving a diagnosis. One SLT shared their perspective:

> But in my experience, 90% of people are in survival mode, before disability even hits your family in [South Africa]. We are all trying to be safe, people are all trying to make money... And then you get a child with a disability that adds so much, that most of us in privileged situations will never understand.

The professionals revealed that sometimes caregivers do not fully understand the implications of their child's diagnosis and "the parents don't come into the schooling environment". SLTs feel that helping caregivers understand the purpose of AAC could improve their perception of AAC use. One SLT said:

> When I say the parents don't get it, I put that on us and myself… because I haven't done the right job. If they're not getting it, it's because I haven't found the right way to help them get it

> or I haven't identified what's important for them and I'm not addressing their needs, which is key.

The participants also mentioned limited professional training as a concern. Some teachers working in LSEN schools are not given enough training at an undergraduate level to educate children with special needs. This is further compounded by that fact that,

> Those kids who look, sound, act more severely disabled, who, for the most part are our AAC kids, are the ones who are left with less trained staff…it's not always the case- it is changing- but that legacy, [still] exists in a lot of our schools.

Furthermore, SLTs mention that AAC resources are often underused in the classroom, "my low-tech stuff is hidden away somewhere." One teacher indicated that "if you want [school-based AAC] to work, you have to invest in the teachers and empower them..." It was discussed that teacher training should focus on giving teachers "the tools" necessary for using AAC in the classroom, rather than training teachers on how to teach the vocabulary. Some teachers have stated that they learn best when SLTs share resources and model how to use AAC. One teacher said, "a lot of their things I imitate, and I ask them to email me this resource and that…"

Similarly, the SLT's revealed that they are also given limited undergraduate AAC training before they are expected to train others to use AAC. One SLT said: **"**You look online and there's lots of lovely, very long, convoluted articles about AAC, that's very helpful, but show me actually how to do it because in [university], I didn't do it." One SLT shared that "it fell on me as a therapist to programme the devices," which wasn't prioritised in undergraduate training. SLTs expressed: "I'm learning [about AAC] as I go" and indicated that they want a "step-by-step guide" that they can work from towards sustainable AAC implementation.

Device and software

The device and software theme encompasses professionals' perceptions of the limitations of high-tech devices in under-resourced settings, with three subthemes: safety, device affordability, and device features.

Professionals mentioned that taking public transport can be dangerous when children have high-tech devices. SLTs also reiterated caregivers' concerns, "if mom says, 'I'm really not comfortable, our home's not safe. Having [a high-tech device] in our home is going to put us at risk'. I don't want that for any of my [clients]. Okay, then the low-tech goes home".

Professionals highlighted device affordability as a barrier to high technology AAC implementation. The professionals would like learners to have assistive devices. However, they reported that schools do not have the financial resources to provide assistive devices, and most of the families "can't afford that technology". One SLT shared their experience, "some of my [clients] use low-tech picture-based systems, but I'm trying to move them on to high-tech, but because they are so expensive, we've had to do a lot of fundraising. So that has been a challenge". Many schools are not comfortable sending high-tech devices home because of the excessive replacement cost. One SLT mentioned: "But the iPad breaks, the charger doesn't come in, we have [extended periods with no electricity], the screen gets smashed, it just freezes…" Access barriers limit AAC use in LMICs, although there are clear benefits for children's communication. For example, some families do not have access to electricity, and professionals do not necessarily have the resources to implement a functional AAC system: "I only used low-tech because there was no high-tech… And we didn't even have internet... So, it was whatever I [could] find at home."

Device features must also be considered in LMICs. One SLT mentioned that "the majority of AAC apps, and particularly the [evidence-based] ones, are actually on Apple, not Android..." but Apple iPads are more expensive than Android devices. AAC software has a limited number of South African voices. Thus, to differentiate the different speakers in the classroom, one needs to "modulate the frequency" of the voices. Speech generation devices, such as Go-Talks by Attainment Company (2011), have been offered to some schools, but consensus was that SLTs did not like these devices. One SLT said:

> I can't stand [Go-Talks]" while another SLT said, "I do have some kids that use Go-Talks. They're not my favourite…for some kids, [Go-Talks] are helpful, but really, it's just about, do the teachers implement it in the classroom? And the answer is largely no.

Education system

This theme considered the challenges that professionals regularly face, working with school-aged children who need AAC to communicate. Three subthemes were identified, including high staff turnover and burnout, therapist-teacher collaboration, and caseload.

According to the SLTs, "there's high staff turnover in schools, you're training all the time". The AAC system that one SLT implements may not be the preference for the next SLT or welcomed by the next teacher. One SLT said,

> It's like pick your battles. The same as you're speaking about teachers coming in, therapists coming in, and the high turnover of staff… the ball almost stops rolling, rolls back a bit and

> then a new therapist spends two years just getting the ball to where you left it. And then she's probably burnt out and needs to take a break. I mean, that ball never makes momentum.

SLTs shared that AAC intervention may not be prioritised because SLTs may feel "uncomfortable with [AAC] systems… [SLTs] just don't know where to start" and new therapists may focus on other areas of communication instead.

Therapist-teacher collaboration was another subtheme mentioned by the professionals. The most common way for therapists to see children during school time is to use the 'pull-out' method (children taken out of class for therapy). One SLT explored some of the disadvantages of this method, "the teacher doesn't get to see [AAC] working. The child doesn't get to see the teacher use it; the teacher doesn't see the therapist use it. The therapist doesn't see what goes on in the class..."

However, another SLT mentioned that sitting in the LSEN class is only possible if "there is space in the classroom". Even so, participants made it clear that the collaboration between therapist and teacher is essential for a positive outcome for the child. Some teachers are accommodating and eager: "a lot of times we [teachers] observe the therapist in the class and then you see certain techniques, certain resources that they use… then you actually learn a lot of the things you can apply with other kids in the classroom", while others stated "you're not sitting in my class, this is my time with my children". Some teachers feel that SLTs make better progress when they see the children outside of the classroom: "[SLTs]… do some magic there [intervention in SLT room] and when they come back, I'm able to teach this child successfully and assess everything." It was evident from the discussion that teachers had different preferences for levels of assistance in the classroom.

Large caseloads were acknowledged as a great challenge for SLTs in LSEN schools. One therapist spoke about her experience:

> We were two SLTs in a school of 200. I had a caseload of about 70, of which at least 10 were children who couldn't speak and needed AAC, but I didn't have enough time to implement systems with them, or train everybody.

SLTs reported that they do not have enough time to conduct intensive therapy with the children, "that means that for that whole week, my other 35 kids don't get seen." SLTs also mentioned that aside from the large caseloads, all their extra time was filled with other school responsibilities, such as "break duty", sport, and "finance committee" work, and they had little time to personalise AAC devices. Additionally, as staff numbers are limited at public schools, one SLT said, "the OTs also have 70-90 [children on their caseload], and there are also wheelchairs that are broken and there isn't someone to fix the wheelchairs. So, then the physios become the wheelchair repair people, or the teachers join". It

was acknowledged that assistants would be valuable for the entire multidisciplinary team, "so that you can take your expertise and maybe manage the 70 [cases] but have multiple assistants on the ground in the classroom". However, school budget constraints meant that assistants are seldom hired.

Language and code-switching

This theme highlights the difficulties experienced by professionals who work with children using AAC in a multilingual country. It has three subthemes, including language, accents, and code-switching.

SLTs shared the language challenges experienced, specifically that there are few synthetic voices in some of the South African languages. In addition, most of these AAC systems are text-to-speech systems, which illiterate children cannot access. One SLT stated,

> Over the years, we never really had a child successfully using a speech output device, largely because of language barriers and accents... They're all isiXhosa speaking children [in the school], and to create an entire vocabulary in a different language, when it's not available, would take years.

This highlights the challenges faced in multilingual contexts. If an AAC system is set up for English, it typically cannot be translated into an African language such as isiXhosa. For example, in isiXhosa, "you can't use one word with a picture, because words aren't words, they're root words." For example, ndiyam**bona** (I see him) is contrasted with bayam**bona** (they see him) and both derive from uku**bona** (to see). The form of the verb changes depending on the context. According to the professionals, some families feel that evidence-based AAC focuses on English and the devices offering additional languages are suboptimal for school. One SLT said, "I've just had [several] Afrikaans families who are choosing English education for their children so that their children can use English AAC devices". Moreover, SLTs must find ways of creating AAC systems in unfamiliar languages and established software does not always support minority languages. One therapist expressed her concern noting, "too many South African therapists are working in AAC in other languages without having proper foundations in those languages." Although SLTs mentioned that there are "researched [evidence-based] core systems" in African languages, these are not freely available. SLTs, therefore, end up making bespoke AAC systems for children in different languages, which is time intensive.

The professionals noted that American, British and Australian English accents are generally used with speech-generating devices. An SLT noted how pronunciation was often problematic for users in South Africa whose English pronunciation of words was very different. Discussing a Zimbabwean child who used a speech generating device, SLTs said,

> He had this American computer voice… I was like, this is such a lovely story, but I cannot focus because it sounds ridiculous … an adult male American voice, but the child had a Zimbabwean name. So even his own [device] couldn't pronounce his own name.

On the other hand, another SLT mentioned: "I've had people who, when the kids speak with an American or… what is to be considered a 'posh' accent, then people actually expect more of them." It seems that when a child uses a high-tech device, the accent of the synthetic speech plays a role in how the child is perceived by others.

Code switching is another aspect of growing up in a multilingual society. Participants believed that almost every family in South Africa alternates between two or more languages during conversation, even if the child has special needs. One SLT described the code-switching situation well by saying:

> I'm not aware of a high-tech device that does [code-switching] in a functional manner, anywhere in the world… most devices still require a significant shut down to get the second language. And so, you can't code-switch, because you can either be in one language or the other… [South Africa's] got such a mix [of languages], that code-switching is a real challenge when it comes to AAC.

Benefits

This theme encompasses the benefits experienced by professionals who work with children that use AAC. It has two subthemes, including attitudes and expectations, and identity and personality. A teacher shared,

> I can see it assists with communication, social interaction and with learning. [Often] you sit in a classroom and a child wouldn't have been able to tell you what they like, what they dislike, their emotions- and I can see the iPad even assists with speech now.

One of the teachers mentioned that having assistive devices for some of the children in their class would reduce the learners' frustration. An SLT summed it up as follows, "in the core of all the… hecticness of trying to get a child to speak, we need to remember the value of that child speaking, of that little personality sharing with you." According to the participants, AAC could change people's attitudes, expectations, and perceptions of children with CCN: "People expect so much less of [children who use AAC] and then you're changing the perception of them, again." Another SLT said: "Having a voice, especially in a classroom can be really great. Because kids who are hidden suddenly are not anymore. Professionals agree that "getting [children with CCN] as independent and as excited as

possible, will give hope to their parents, will give hope to the teacher, will give hope to the therapist to push a little bit harder."

The professionals expressed how children can share more of their personality and identity when they are given a voice:

> If this voice does give them their identity, their excitement for sharing, for communicating, I mean, the world's their oyster. In comparison to this child that may be like- 'what I have to say, doesn't really matter, because what's going to come out? That's not me'.

As age plays a role in identity, SLTs shared that having a voice that ages with the child is an equally important identity component to consider but, "sometimes we neglect [aging the voice] for the… population." Lastly, the professionals acknowledged that AAC has the potential to give children the ability to share their thoughts, helping them feel less isolated. One SLT shared, "the discussions I've had with young teens to adults, who have no speech and who have to use these voices has been very much that when they find their voice, it's like that moment of finding who they are".

Discussion

Literature acknowledges that communication partners are crucial to the successful use and implementation of AAC systems (Jette et al., 2017; Kent-Walsh et al., 2015; Moorcroft et al., 2019b). However, published literature from AAC stakeholders is not representative of lows. For example, in Baxter et al.'s (2012) systematic review, stakeholders from 27 studies resided in high-income countries such as the UK, USA, Australia, Canada and Israel, but only one study included perspectives from a middle-income country, Egypt. Although Baxter et al.'s (2012) research was conducted over 10 years ago, recent literature focusing on LMICs is still less common than research from high-income countries, and often lacks input from multi-perspective stakeholders (Mukhopadhyay and Nwaogu, 2009; Tönsing et al., 2019; Tönsing and Dada, 2016; van Niekerk et al., 2019). This exploratory study presented perspectives from teachers, SLTs and caregivers in South Africa, an LMIC. This study was conducted to highlight additional issues for AAC implementation in this context as little information is available, but further work will be needed before generalisation can occur.

Findings indicate that reduced support and training is the most commonly experienced challenge that professionals face when implementing AAC. Although a lack of support and training is not unique to LMICs (Moorcroft et al., 2019a), when combined with limited personnel, training and resources (Pascoe and Norman, 2011), the problem is exacerbated. Our study highlights that without adequate support and training, AAC service delivery in LMICs may remain unsustainable. Other South African studies (Tönsing et al., 2019; Tönsing and Dada, 2016; van Niekerk et al., 2019) echo the need

for training when introducing assistive products. Similarly, results from our caregiver focus group suggest that misunderstandings surrounding the purpose of AAC are experienced with caregivers, often brought about by a lack of AAC-related training and education, as supported by van Niekerk et al. (2019). Training caregivers from LMICs is challenging as they often cannot attend meetings due to a lack of transport and limited time off work. They may also have limited access to technology and/or basic resources. This was clear in our study and research from other developing countries, such as Kenya (Gona et al., 2014). Professionals highlighted the need for therapist and teacher empowerment to implement change, and support from the education system. Providing training and support to stakeholders so that they may be agents of change for AAC provision and intervention could pivot the burden of disease for children who need AAC, and their families.

Although South Africa has a progressive disability policy, the accessibility of assistive devices, particularly high-tech AAC devices, is limited (van Niekerk et al., 2019). When high-tech devices are available, our study shows that practitioners from LMICs need to consider the high risk of crime and the risk that using such an expensive device places on the child and their family. Communication difficulties, along with the attached stigma and cultural beliefs surrounding children with special needs in LMICs, may act as a catalyst for bullying and crime. These findings are supported by studies in other LMICs (Gona et al., 2014; Mukhopadhyay and Nwaogu, 2009). However, even if the family's safety is not at risk from having a high-tech device in the home, many families, even those in high-income countries, prefer that their children only use high-tech devices at school (Tegler et al., 2019). If the language used at school differs from the language used at home, high-tech devices become less useful (Tönsing et al., 2018). It is recommended that children receive a low-tech device that matches the layout of the high-tech device. This low-tech device could be used in situations where having a high-tech system puts the family at risk, as a back-up in case the high-tech system fails, or in circumstances where the high-tech system is impractical or unfeasible.

Additionally, our study highlights that practitioners need to consider high-tech AAC device affordability and device features. Although much of Africa's AAC development work has been conducted in South Africa (Gona et al., 2014), the country is limited by financial resource constraints (Tönsing and Dada, 2016; van Niekerk et al., 2019; Visagie et al., 2020). There is great variation in resourcing and the provision of assistive technology between the different provinces (regions), with more rural provinces often experiencing the greatest challenges (van Niekerk et al., 2019; Visagie et al., 2020). People who provide and use assistive technology in low-resourced contexts often support imported products over local ones (Visagie et al., 2020). Imported products are usually more expensive, and the design features may not be suited to the environmental, cultural and language needs of the country (Visagie et al., 2020). If local companies were incentivised with governmental tenders, production costs may decrease, and the availability of affordable local technology and culturally

appropriate software would increase. It would be beneficial if collaborative and strategic partnerships and communication channels between the Departments of Health and Education, non-governmental organisations, disabled persons' organisations, local manufacturers, and donors, were created and formalised to ensure affordable and accessible AAC solutions.

Local research into culturally appropriate synthetic voices and other AAC systems should be ongoing, and current best practice should not simply rely on research from high-income countries. Our study showed that practitioners need to consider the range of languages and accents available when introducing AAC systems. Many LMICs are multilingual, and may lack access to minority languages on their AAC devices (Terblanche et al., 2022). Moreover, professionals are not always sufficiently competent in all the languages to meet the language demands of a multilingual society (Kathard et al., 2011). Despite ethical guidelines stating that individuals should not be denied intervention because of a language mismatch (Pascoe and Norman, 2011), our study shows that evidence-based core language AAC systems are not widely available for SLTs and teachers, especially when finances are limited. Simple translation of a graphic symbol-based system into two or more spoken languages is not appropriate because different languages require different representation methods (Tönsing et al., 2018). However, even if systems were developed, they would need to allow code-switching, and make use of culturally and linguistically appropriate AAC symbols, paired with the correct orthography (Tönsing et al., 2018). We need to advocate for government-supplied culturally appropriate core language AAC systems. In multilingual countries, the cross-cultural readability of graphic symbols, and the development of AAC systems in different African languages, would assist AAC users and those who support them.

The development of culturally appropriate core language AAC systems may ease the burden on professionals. Large caseloads, a lack of human resources and related support services in LMICs, mean that professionals in LSEN schools feel overworked, leading to periods of burnout. Similar experiences were shared by professionals in Botswana (Mukhopadhyay and Nwaogu, 2009) and Egypt (Wormnaes and Malek, 2004). As assistants are not always available in LMICs, SLTs report that classroom-based intervention for each child on their caseload is not necessarily practical. Our study highlights that if LSEN schools do not prioritise and advocate for AAC use, AAC development for children with CCN will remain limited.

Despite these challenges, there are numerous benefits to using AAC systems. AAC gives children with CCN a functional way to initiate and respond to conversation, which reduces social isolation and encourages active participation. Children with CCN have opportunities to showcase their personality, interests, and ideas. When using a high-tech AAC device, having a synthetic voice that mirrors the language, gender, age, and natural voice capabilities of a child with CCN, provides that

child with an identity, and strengthens their communication partners' expectations and perceptions of them.

Limitations and Future Directions

This is an exploratory study and therefore the small sample size and methodology selected means that the findings cannot be generalised to all AAC stakeholders in LMICs. The caregiver group was particularly limited. Although the focus groups were conducted after working hours and transport fees were covered by the study, the caregivers informed the research team that the loss of income, the transport cost which had to be paid upfront, and lack of childcare for the other children in the home, were all reasons that made participation difficult. These challenges indicate clearly why similar participants from low-income backgrounds often go underrepresented in research. Future research should consider gathering the views of stakeholders at other ecological levels and mapping the process of change through theory of change workshops. Users of AAC themselves should also be included as stakeholders in future projects. Researchers need to develop multilingual AAC systems and AAC intervention guides, along with appropriate synthetic child speech in under-resourced languages.

Conclusion

Communication partners are vital to the successful use and implementation of AAC systems. This study presented the perspectives of SLTs, teachers and caregivers of children using AAC in South Africa, highlighting practical issues professionals need to consider when implementing AAC in under-resourced contexts. The safety risk associated with using high-tech AAC devices, device affordability, device features, and the range of languages and accents available when introducing an AAC system are all aspects that are important to consider in this context. Strategies that may advance the implementation and use of AAC in South Africa, and other similar settings, include development of strategic partnerships between governmental and non-governmental groups, establishing appropriate communication, training, and support systems, and creating evidence-based core-language AAC systems.

MANUSCRIPT THREE

The development of synthetic child speech in three South African languages.

Camryn Terblanche[1], Tyler T Schnoor[2], Michal Harty[1] and Benjamin V Tucker[3,2]

[1]Division of Communication Sciences and Disorders, University of Cape Town, SA

[2]Department of Linguistics, University of Alberta, Canada

[3]Department of Communication Sciences and Disorders, Northern Arizona University, US

Abstract

It is well-known that children with expressive communication difficulties have the right to communicate, but they should also have the right to do so in whichever language they choose, with a voice that closely matches their age, gender, and dialect. This study aimed to develop naturalistic synthetic child speech, matching the vocal identity of three children with expressive communication difficulties, using Tacotron 2, for three under-resourced South African languages, namely South African English (SAE), Afrikaans, and isiXhosa. Due to the scarcity of child speech corpora, 2 hours of child speech data per child was collected from three 11- to 12-year-old children. Two adult models were used to "warm start" the child speech synthesis. To determine the naturalness of the synthetic voices, 124 listeners participated in a mean opinion score survey (Likert Score) and optionally gave qualitative feedback. Despite limited training data used in this study, we successfully developed a synthesized child voice of adequate quality in each language. This study highlights that with recent technological advancements, it is possible to develop synthetic child speech that matches the vocal identity of a child with expressive communication difficulties in different under-resourced languages.

Keywords: augmentative and alternative communication (AAC), children, expressive communication difficulties, neural network, South Africa (SA), speech synthesis, Tacotron 2, text-to-speech, under-resourced languages

Introduction

Research in augmentative and alternative communication (AAC) has often discussed an individual's right to communicate, but the rights of individuals using AAC to communicate in whichever language they choose, particularly languages spoken in low- and middle-income countries, has not been equally prioritized (Tönsing et al., 2019). High-income, English-speaking countries often dominate AAC research and development (Tönsing et al., 2019). As a result, American English is most commonly found on speech-generating devices (Terblanche et al., 2022). Languages and dialects from low- and middle-income countries, such as South Africa, are often underrepresented (Sefara et al., 2019). This means that the speech output available on high-tech AAC devices, given to children with expressive communication difficulties to assist with functional communication in low- and-middle income countries, is often not reflective of the child's natural voice, and there is regularly an age, gender, language and/or dialectal-mismatch on these devices (Jreige et al., 2009; Mills et al., 2014). Although South Africa is well-known for its multilingualism, as there are 12 official languages, including sign language, and many unofficial languages and dialects, this linguistic and cultural diversity is not well-represented on South African AAC devices (Sefara et al., 2019), as many of these official languages are considered under-resourced. Languages are considered under-resourced when they have a limited online presence, there is a lack of linguistic expertise on the language, there are limited data for speech and language processing, reduced transcribed speech corpora and pronunciation dictionaries, as well as limited resources for speech, language, and literacy development (Besacier et al., 2014). Despite under-resourced languages having limiting expertise and resources, they are not necessarily the minority languages in the country (Besacier et al., 2014), and there is often a substantial user base. For instance, according to Statistics SA (2011), more South Africans speak Afrikaans (6.8 million), isiXhosa (8.15 million) and isiZulu (11.58 million) as their home language than South African English (4.89 million).

Limited access to devices that incorporate their home languages may further marginalize children who rely on AAC to communicate. Children using a speech-generating device in low- and middle-income countries are likely using their second or third language to communicate (Tönsing et al., 2019), with many sharing the same synthetic voice (e.g., adult US-English male) as all the other children with expressive communication difficulties in the classroom (Mills et al., 2014). Due to the dearth of contextually relevant resources in low- and middle-income countries, such as South Africa (Pascoe & Norman, 2011), the lack of African AAC resources, including applicable child voices in African languages, results in African AAC users being underserved. Creating linguistically and culturally appropriate synthetic speech in low- and middle-income countries begins to address the historical linguistic imbalance and discrimination imposed onto under-resourced and indigenous languages (Sefara et al., 2019). In Tönsing et al.'s (2019) study, South African participants with

expressive communication difficulties shared that they could not express themselves in all the languages that they understood and were exposed to and were mostly limited to English. These participants viewed their ability to express themselves in certain languages as part of their identity. They therefore expressed a desire for the development of appropriate AAC systems and interventions in different languages, as well as increased literacy learning opportunities (Tönsing et al., 2019). In addition, it has been found that using a specific language in a multilingual setting allows one to show respect, promote group cohesion and align or distance oneself from communication partners (Ndlangamandla, 2010). According to the South African Constitution (South African Government, 1996), every citizen has the right to use whichever language they choose. This basic right should not be overlooked for individuals with expressive communication difficulties who are not given access to speech generating devices that feature their home language. Until AAC applications can meet the communication needs of children with expressive communication difficulties, AAC use, particularly in low- and middle-income countries, will likely remain limited, resulting in fewer opportunities for participation, interaction, and academic growth for individuals with expressive communication difficulties.

As an individual's voice is an indicator of their personality, age, sex, social background, and cultural identity (Jreige et al., 2009; Mills et al., 2014; Sutton et al., 2019), the voice used on speech generating devices for children, should, if possible, at least approximate the child's voice and linguistic background. The development of child voices in under-resourced languages is therefore an important task. However, due to the scarcity of child speech corpora, generating natural child speech synthesis is a difficult task. There are several challenges associated with the collection of child speech data for the creation of a synthetic voice. For example, children's read speech is generally less fluent as compared to adult speech, particularly if children have reduced literacy skills in their home languages. Children's speech includes articulatory inaccuracies, and often presents with typical disfluencies and hesitations. Moreover, child speech recordings are typically conducted in insufficient recording environments, and as a result background noise is a common occurrence (Govender et al., 2015). However, in instances when the available data is limited, implementing a warm start has been proven useful (Phuong et al., 2021). Essentially, transfer learning is a specific example of warm starting, where the weights of one model are used as the initialization point for training another model, i.e., established synthetic speech models act as a skeleton for newer models. In other words, using more widely available adult data as a skeleton to create a child's voice. There are documented benefits to implementing a warm start. First, using a warm start can greatly reduce the model's training time (Barnekow et al., 2021; Cooper et al., 2020; Zhu, 2020). Second, using a pre-trained English model to warm start models in other languages, has proven to be very effective (Phuong et al., 2021; Saam & Cabral, 2021; Zhu, 2020). This is due to the fact that the pre-existing model's voice is erased, but the speaker-general characteristics, which can be transferred to other speakers and languages, remain intact, thereby reducing the overall iteration and/or training time of the new model (Barnekow et al., 2021). For example, Phuong et al., (2021) used

an English pre-trained model with 1 hour of Vietnamese data and found that the quality of the resultant speech output was as good as training a model without a warm start, with 5 hr of Vietnamese data. When child speech data is available, there are various freely available speech synthesis systems that can be used to develop synthetic voices. One open-source speech synthesis system that produces synthetic speech with high naturalness and good perceptual similarity to target speakers/donor voices is known as Tacotron 2 (Shen et al., 2018; Wang et al., 2020). Tacotron 2 is an end-to-end neural network-based text-to-speech (TTS) system that can be trained on text-to-audio pairs, without substantial phonetic annotation. These technical advances mean that data from well-resourced languages can now be recruited much more effectively than previously possible to develop synthetic voices for under-resourced languages.

Due to these recent advancements in accessible text-to-speech technology, this study aims to accurately develop synthetic child speech in three different under-resourced languages that match the vocal identity of three target children with expressive communication difficulties. This paper is therefore primarily a description of practice and intervention development. First, we set out to determine if it is feasible to develop naturalistic child speech synthesis for three under-resourced South African languages, namely Afrikaans, isiXhosa, and South African English (SAE), using Tacotron 2, an open-source speech synthesis system. Second, we solicited feedback from first-language speakers in terms of the naturalness of the synthetic voices we developed.

Method

This project involved two phases. In Phase 1, child speech data for each of the three languages was collected. Following this, we used open-source adult speech corpora to develop synthetic adult speech in SAE, Afrikaans, and isiXhosa (Louw & Schlünz, 2016a, 2016b, 2016c). We incorporated these adult models, along with the collected child speech data, to develop child speech synthesis in each language. The synthetic child speech was created for three specific children with expressive communication difficulties who attend different schools for learners with special education needs, in Cape Town. In Phase 2, we evaluated the synthetic speech. The subjective mean opinion score (MOS) method was used to gather listeners' perceptions of the naturalness of the synthetic voices. Additionally, participants were able to optionally provide a qualitative description of their impression of the synthetic speech in response to an open-ended question, "Is there anything else you wish to share about the voices?"

Phase 1: Development of the Synthetic Child Speech

Participants

Spontaneous (5 min) and read speech (5 min) samples from 98 children who were typically developing were obtained in a classroom repurposed as a storeroom at one mainstream English school in Cape Town. The repurposed classroom had carpets, few windows, and the additional books and furniture were helpful to reduce reverberation and background noise. Based on the similarity of the children who were typically developing to the three children with expressive communication difficulties, for whom the synthetic child speech was intended (i.e., age, gender, home language, demographic group), as well as their performance in the picture description and reading tasks, only three of those 98 children who were typically developing were selected to further participate in the study. Two 12-year-old children who were typically developing (one male and one female) and one 11-year-old male child who was typically developing were recruited to each record a total of 2 hr of read speech. A purposive sampling method was used to collect the speech samples from the children. In this method, the researcher intentionally selects participants based on their particular qualities, specifically their similarity to the three children with expressive communication difficulties for whom the voices were being developed. Although purposive sampling is subjective, Taherdoost (2016) states that this method is convenient, low cost, less time consuming and ideal for exploratory research. All required approvals for collecting the children's speech data were obtained; guidelines outlined in the Helsinki Declaration of 2013 were followed (World Medical Association, 2013). The provincial education department and the relevant school principals gave permission to access the schools. All the children and their caregivers gave informed consent to audio record and use the speech data for this research.

Materials and Measures

To select suitable child participants who were typically developing, a picture description task (The School Aged Language Assessment Measures/SLAM) was used. SLAM's Ball Mystery Story (Crowley, 2019) includes picture cards and questions to elicit connected speech for children aged between 9-15 years old. In addition, various age-appropriate library books were offered to the children for the reading task, and based on their interests, the children selected their preferred books to read in their home languages. A Zoom H1 Handy recorder [3] was used to collect the children's speech data.

As the child speech data was limited, this study also included various freely available adult models and corpora. First, the creators of Tacotron 2 published a pre-trained Tacotron 2 model from NVIDIA[1]. This established model was used in this study. The LJ dataset (Ito & Johnson, 2017) was used to train the Tacotron 2 model, and it is extensive, with *approximately 24 hr of* short recordings from one female adult North American English (NAE) speaker (Ito & Johnson, 2017). Second, the

Afrikaans (Louw & Schlünz, 2016a), isiXhosa, (Louw & Schlünz, 2016c) and the SAE Lwazi III text-to-speech datasets (Louw & Schlünz, 2016b) were also used. Each Lwazi III dataset is made up of short and long audio clips from one female adult speaker per language. Each language has approximately 6-7 hr of data. Although each Lwazi III dataset contains mostly one language, there are also small amounts of additional language data, which provides phone coverage of other languages (Louw & Schlünz, 2016b). For example, the isiXhosa Lwazi III dataset contains 5 hr and 53 min of isiXhosa but also includes SAE (32 min), Afrikaans (3 min), isiZulu (5 min), Sependi (7 min), and Setswana (4 min) data. It was hypothesized that the inclusion of additional phone coverage might improve the model's pronunciation of non-English names and surnames. It was also suspected that the additional language data may improve the pronunciation of loan words, which are frequently exchanged between languages. To develop the synthetic adult voices, Tacotron 2 (Wang et al., 2017) was used. Tacotron 2 is a state-of-the-art, open-source speech synthesis system that generates synthesized speech directly from graphemes and consists of a recurrent sequence-to-sequence mel-spectrogram prediction network (Wang et al., 2017).

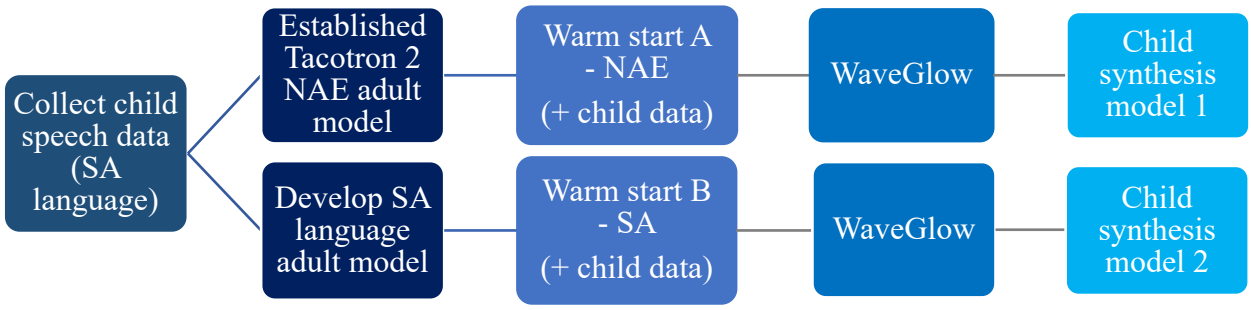
The published pre-trained Tacotron 2 model from NVIDIA[1] was one of the models used to warm start the child speech synthesis (warm start A). The pretrained adult models in each language were also used as an alternative warm start for the child speech synthesis (warm start B). Tacotron 2 (Wang et al., 2017) was also used to develop the synthetic child speech, and a Titan V GPU from NVIDIA[2] was used for all training.

Procedures

The synthetic child voices were developed in three stages. Figure 1 gives an outline of the process used to generate the synthetic child speech in Part 1. First, the speech data from three children who were typically developing was collected (the left-most component in Figure 1). The three children who were typically developing were asked to read in their home languages, namely Afrikaans, isiXhosa, and SAE. Recording sessions lasted 30 min, and the children were given breaks of approximately 5 min every 10-15 min. Using a Zoom H1 Handy recorder (44100 Hz, 16 bit)[3], the recordings were collected in a repurposed classroom. Tacotron 2 has a shorter training time when audio files are ≤13 sec. Thus, the data were segmented into short utterance chunks of ≤13 sec. As fluent speech is preferred for speech synthesis, all false starts, disfluencies, and misarticulations were manually removed from the data. After data cleaning, *approximately* 15-30 min of data per child were lost, leaving 108.7 min for Afrikaans, 101.9 min for isiXhosa, and 113.7 min for SAE. The sound files were split by manually marking utterances using Praat TextGrids. Individual single-channel files were extracted from each of the marked-up utterances and downsampled to 22050 Hz. The data were then randomly divided into training (90%) and validation files (10%).

Figure 1

Overview of the Process to Generate Synthetic Child Speech for Each South African Language Using Tacotron 2



Second, as adult models were required to warm start the child speech synthesis, adult models first had to be developed (the second component in Figure 1). Using Tacotron 2's (Wang et al., 2017) default architecture, the three adult models were trained on cleaned, resampled (22050 Hz) Afrikaans (Louw & Schlünz, 2016a), isiXhosa, (Louw & Schlünz, 2016c) and SAE Lwazi III text-to-speech datasets (Louw & Schlünz, 2016b) (see supplementary Table 1 for further information). Due to Lwazi III's relatively restricted datasets, a warm start training procedure was implemented for all three languages, using the published pre-trained Tacotron 2 model from NVIDIA[1], trained on the LJ dataset (Ito & Johnson, 2017). As transfer learning allows the speaker-general characteristics from the pre-existing model to be transferred to other speakers and languages, the model's training time is reduced (Barnekow et al., 2021; Cooper et al., 2020). This technique is useful when there is limited data, such as with under-resourced languages. During training, each language's full Lwazi III dataset, including the additional language data, were used. After minimal cleaning and segmenting of the Afrikaans and English datasets, training took approximately 5 days per language. However, the isiXhosa audio files were longer and therefore, the batch settings had to be significantly reduced to reduce GPU memory constraints, which resulted in an increased training time of 13 days. Using the generated mel-scale spectrograms from Tacotron 2, a pretrained, publicly available WaveGlow (Prenger et al., 2018) model was used as a vocoder to synthesize time-domain waveforms. WaveGlow is a flow-based generative speech synthesis program (Prenger et al., 2018). Some undesirable artefacts and unwanted background noise were removed by using the denoising code from the WaveGlow repository.

Third, due to the limited child speech data collected, two distinct warm start models, using the adult speech data, were used during the creation of the child speech synthesis (the third component in Figure 1). In warm start A, the child models were trained on the published pre-trained Tacotron 2 model

from NVIDIA[1]. For warm start B, each child model was trained on the respective South African adult model, using the Lwazi III datasets (see supplementary Table 2 for further information). Training for the child speech synthesis took approximately 72 hr for each language, no matter the warm start training procedure utilized. WaveGlow (Prenger et al., 2018) was used as the vocoder to synthesize time-domain waveforms and the denoising code from the WaveGlow repository was implemented (the fourth component in Figure 1).

Phase 2: Evaluation of the Synthetic Speech Via Listener Perception Tests

Participants

In Phase 2, 124 South African participants who spoke the language/s rated the naturalness of the synthetic voices. Snowball sampling was used as colleagues and friends were initially recruited via social media to give their opinions on the voices, and they were encouraged to share the anonymous MOS survey link with others. The survey was therefore shared nationwide, and any South Africans over the age of 18 were eligible to participate. One participant was excluded as they were not South African, leaving 123 participants. The age range of the participants was 18-71 years old ($\bar{x}$= 32 years old). There were 103 females, 19 males, and 1 participant who identified as 'other'. Participants indicated which language/s they spoke and were only presented with stimuli in all the languages applicable to them. Overall, there were 60 Afrikaans listeners, 111 South African? English listeners and 19 isiXhosa listeners who participated (some participants spoke more than one language). As Phase 2 involved an online MOS survey, participants conducted the survey from the comfort of their homes or places of work.

Materials and Measures

In Phase 2, jsPsych (de Leeuw et al., 2012), a JavaScript framework for creating behavioral experiments that run in a web browser, was used to run the online MOS survey. MOS is often used to judge the quality of synthetic voices (Govender & de Wet, 2016; Jain et al., 2022).

Research Design

In Phase 2, a non-experimental quantitative descriptive design was used to evaluate the naturalness of the synthetic voices using MOS. With a quantitative descriptive design, researchers aim to systematically investigate phenomenon through the collection of numerical data, obtained either via surveys or observation, and undertake statistical techniques to analyze the data (Mertler, 2016). All required approvals were received from the appropriate ethics committee; guidelines outlined in the

Helsinki Declaration of 2013 were followed (World Medical Association, 2013). Each participant provided informed consent before beginning the MOS survey.

Procedures

In Phase 2, participants were asked to rate the naturalness of the synthetic adult and child speech using a 5-point Likert Scale, from 0 (*completely unnatural*) to 4 (*completely natural*). Speech naturalness can be described as how well the speech matches a listener's standards of rate, rhythm, intonation, and stress, hence determining how natural the synthetic speech sounds as compared to spoken speech (Sefara et al., 2019). In addition, at the end of the MOS survey, participants were able to optionally provide a qualitative description of their impression of the synthetic speech. The online MOS survey was open for three weeks and participants listened to both synthetic child and adult speech. Listeners were asked to listen to audio clips of words and semantically predictable sentences of differing syllable lengths. For instance, the audio was regarded as short for words of approximately three syllables. The audio was regarded as medium length when sentences had approximately eight syllables, while audio was regarded as long when sentences had approximately 15 to 23 syllables. Research on child speech synthesis has shown that children's speech is difficult to interpret using semantically unpredictable sentences, which is generally used during conventional listening tests (Govender & de Wet, 2016). Therefore, semantically predictable sentences were used during these listening tests. Words and sentences were randomly presented to the listeners and participants were able to listen to the audio as often as they liked before moving on to the next audio clip. Participants first rated the child voice and then the adult voice in one language, before moving on to the next language, if applicable.

Each participant made 18 MOS ratings per language: five MOS ratings for the synthetic child speech implementing warm start A, five MOS ratings for the synthetic child speech implementing warm start B, and eight MOS ratings for the synthetic adult speech. There were 1080 Afrikaans responses, 1998 English responses, and 342 isiXhosa responses. Using R (R Core Team, 2019), both descriptive and inferential statistics were conducted on the MOS data. Considering the inferential statistics, we use an ordinal mixed-effects regression model. Individual Likert-type questions are generally considered ordinal data because the items have clear rank order, but don't have an even distribution, and a mixed effects model is a statistical test used to predict a single variable using two or more other variables. It is also used to determine the numerical relationship between one variable and others. The predictors for the ordinal mixed-effects regression model are overall ratings, the speaker (adult vs. child speech), the language (SAE vs. Afrikaans vs. isiXhosa), the warm start type (warm start A vs warm start B), as well as the amount of child speech training data required for improved quality. The model also makes use of random effects, which include the participants and the audio stimuli.

Results

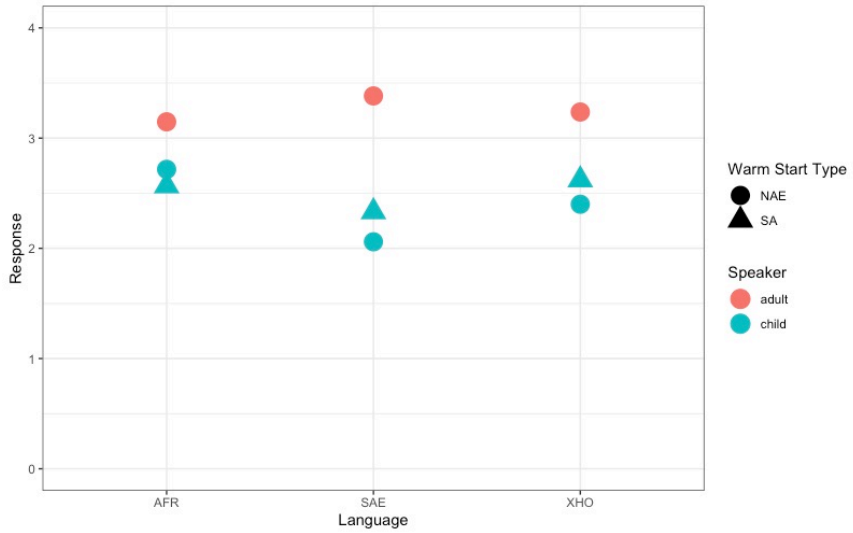
The results section considers the MOS results and the qualitative feedback that some of the listeners provided. This section compares the overall ratings, the speaker, the language, the warm start type, as well as the amount of child speech training data required.

Overall Ratings

Encouragingly, we were able to develop both a synthesized child voice and a synthesized adult voice, in each language. It is important to note that carefully cleaned speech data was preferrable over dirty speech data (e.g., data with false starts). For instance, the synthetic child voices with clean training data were impressionistically less noisy, with improved accuracy. Although the child speech was trained over a different adult model, the accent, language, and vocal acoustics are comparable to the South African child donor, rather than the adult voice. One participant said, "The [English] child's voice was clearly a Western Cape accent. Initially difficult to understand (culturally) but then got more used to it and it sounded natural".

Figure 2

MOS Responses, with Reference to Speaker, Language, and Warm Start Type



Note. This figure presents MOS responses categorized by speaker type (adult and child), language (Afrikaans/AFR, South African English/SAE, and isiXhosa/XHO), and warm start type (North American English/NAE, and South African/SA). The vertical axis represents MOS responses from 0 (completely unnatural) to 4 (completely natural), while the horizontal axis denotes the different languages. Each language corresponds to a specific combination of speaker and warm start type, providing a comprehensive overview of the perceived speech synthesis quality across these dimensions.

The MOS results gave further insight into the naturalness of the voices. The average naturalness rating across all voices was 2.72. Figure 2 illustrates the average MOS, categorized by speaker, language, and warm start type. To test the statistically significant effects, an ordinal mixed-effects regression model was implemented with the ordinal package (Christensen, 2022) in R (R Core Team, 2019), with the results presented in Tables 1 and 2.

Table 1

Fixed Effects Coefficients of All the Voices

Fixed effects	Estimate	*SE*	*z*	*p*
SAE language	-0.36	0.27	-1.32	.188
isiXhosa language	-0.49	0.31	-1.54	.123
Child speaker	-1.98	0.23	-8.49	<.0001
Syllables	0.01	0.02	0.37	.72

Note. Statistically significant coefficients are denoted by p-values less than .01.

Table 2

Fixed Effects Coefficients of the Child Voices

Fixed effects	Estimate	*SE*	*z*	*p*
SAE language	-1.02	0.41	-2.48	< .01
isiXhosa language	-0.62	0.46	-1.36	.17
Warm start B	0.20	0.34	0.59	.55

Note. Statistically significant coefficients are denoted by p-values less than .01.

Speaker

The average natural rating across all the adult voices was 3.25, while the average natural rating across all the child voices was 2.45. An ordinal mixed-effects regression model predicting mean opinion score as a function of language, speaker, and syllable length, was run (formula: Response ~ Language + Speaker + Syllables). The model's intercept corresponds to Language= Afrikaans, Speaker= Adult. Table 1 highlights the fixed effects coefficients for all the voices. Table 1 displays the estimated coefficients, standard errors (SE), z-scores (z), and p-values (p) for each fixed effect in the regression model. The coefficients represent the estimated change in the dependent variable associated with each predictor variable while holding other variables constant. Statistically significant coefficients are denoted by p-values less than .01.

There was a statistically significant effect between adult and child voices, with the adult voices consistently rated as more natural as compared to the child voices. This effect is illustrated in Figure 2, as the MOS responses corresponding to the adult voices are higher than the child voices, indicating greater naturalness ratings for the adult voices. Moreover, one participant expressed their opinion well by saying, "I feel that the adult voices are definitely more clear, and natural, but I really enjoy the children's voices, especially if it's for a child- it makes them feel like they can be who they are supposed to be!*"*

Additionally, Table 1 shows that there is not a statistically significant difference in naturalness between synthetic sentences of differing syllable lengths for either the child or adult voices. Similarly to findings from Jain et al. (2022), the first and last words in the synthetic child speech were more likely subject to distortions and artifacts, as compared to the middle of the phrase. As typical children's speech is less fluent when compared to adults (Jain et al., 2022), any articulatory inaccuracies consistently produced, were also present in the synthetic child voice. This was observed in all three languages. For example, in SAE, the child donor had a non-standard pronunciation of /ld/, such as in wor**ld** and chi**ld** and in isiXhosa, the child donor simplified the isiXhosa click sound /!/, used in words such as ug**q**irha (doctor) and umn**q**wazi (hat). This inconsistency was noted by one isiXhosa listener who said that "[the isiXhosa child] is not pronouncing the clicks correctly," Conversely, audio of the donor's breathing remained in the training data, which transferred to the synthetic models, making them appear more realistic, according to listeners.

Language

In Figure 2, we can see that the Afrikaans adult voice was rated least natural, while the Afrikaans child voice was rated as most natural. Conversely, the English adult voice was rated as most natural, while the English child voice was rated as least natural. No statistically significant effect was

found between any of the adult voices. It should be noted that the additional mixed-language data used in the adult synthesis did not improve the pronunciation of non-English names and loan words to a great degree. Regarding the synthetic child voices, a model of mean opinion score as predicted by language and warm start type is shown in Table 2 (formula: Response ~ Language + Start). The model's intercept corresponds to Language= Afrikaans child, Start= Warm start A. Table 2 displays the estimated coefficients, standard errors (SE), z-scores (z), and p-values (p) for each fixed effect in the regression model related to the child voices. The coefficients represent the estimated change in the dependent variable associated with each predictor variable while holding other variables constant. Statistically significant coefficients are denoted by p-values less than .01.

There was a statistically significant effect between the Afrikaans child voice and the SAE child voice. The SAE child voice was considered less natural than the Afrikaans and isiXhosa child voices by listeners. Additionally, as the isiXhosa child donor's language of learning and teaching was English, her isiXhosa literacy skills were reduced in comparison to her English literacy skills, which resulted in a slower reading rate, thereby affecting the pacing of the isiXhosa synthetic child speech. This was also observed by an isiXhosa participant who shared that, "[The isiXhosa child] sounds like it's an American learning to speak isiXhosa, it was like listening to the audio of Wakanda. [The isiXhosa adult] is audible and sounds natural."

Warm Start Type

Table 2 also highlights that child voices created by implementing warm start B (South African/SA adult voices) were not significantly different from warm start A (North American English/NAE voice). Although not statistically significant, Figure 2 suggests that listeners appear to prefer the SAE and isiXhosa child models when warm start B (SA speaker) was used. In Figure 2, MOS responses for warm start B are higher than warm start A for the SAE and isiXhosa child voices, indicating greater naturalness for warm start B. The opposite can be observed when looking at the Afrikaans child voice, as Afrikaans listeners appear to prefer the Afrikaans model when warm start A (NAE speaker) was used. This preferential pattern is also observed when comparing the consistency of each model's speech output (see supplementary Figure 1). It appears that the SAE and isiXhosa child models are more consistent when warm start B (SA speaker) is used, while the Afrikaans child model is more consistent with warm start A (NAE speaker).

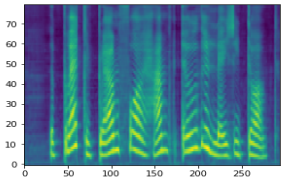
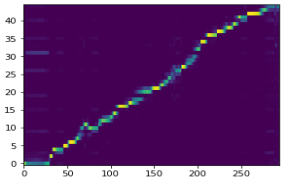
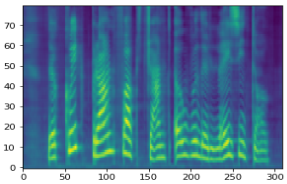
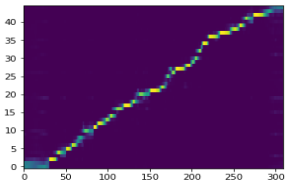
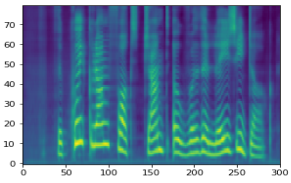
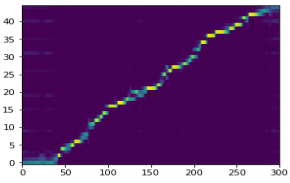
Child Speech Training Data

Using the SAE child speech data, six models with increasing training length were run to determine how much data would be needed to develop child speech synthesis in SAE. Warm start B was used for each model, and models were run with 5 min, 10 min, 20 min, 30 min, 60 min, and the

full 113 min of SAE child speech data. Figure 3 depicts the mel-spectrograms (Figure A) and alignment plots (Figure B) for some of the SAE child models.

Figure 3

Tacotron 2 Mel-Spectrogram (a) and Alignment (b) Plots of Synthesized Speech: "The Quick Brown Fox Jumped Over the Lazy Dog"

	Warm start with SAE	
	Figure A	Figure B
Model 1: 4.82 min		
Model 5: 60.26 min		
Model 6: 113.70 min		

Note. The mel-spectrogram is a spectrogram with the mel scale as its y-axis. It is a good indicator of the signal strength at various frequencies in the waveform. The alignment plot is a quick way to visualize a model's success. A straight diagonal line from the bottom left to the top right is a good indicator that the model is producing something similar to speech.

Mel-spectrograms (Figure A) allow us to examine how accurately the model captures the nuances of the child donor's speech, such as higher pitch, shorter durations, and different phonetic patterns compared to adult speech. In contrast to Model 6, which displays clear and well-defined spectral patterns, Model 1 shows less distinct spectral characteristics due to the presence of additional noise in the synthetic speech. Moreover, each point on the alignment plot (Figure B) represents a pair of corresponding elements, i.e., the synthesized speech and the donor speech. Figure B shows that our child synthesis model aligns well with the donor speech with regards to intonation, rhythm, and specific phonetic features, which is particularly observed in Model 6. Based on impressionistic listening, the mel-spectrograms and alignment plots, as little as 5 min of training data can produce a speech-like output. However, the speech output for 5-10 min of training data did not consistently match the text, which is not ideal for a text-to-speech voice in the real world. As expected, more training data improves

the accuracy and the quality of the child speech output, as real gains were perceptually achieved after 60 min of training data. Despite the occasional pronunciation and prosodic irregularity, adequate quality child speech synthesis was developed with fairly little child speech data (only 113 min), using pre-existing models (adult-child), to warm start the training.

Discussion

This study addressed the feasibility of creating synthetic child speech for children with expressive communication difficulties, using a limited training data scenario for under-resourced languages, namely Afrikaans, isiXhosa and South African English/SAE. Less than 2 hr of child speech data was used with adult speech corpora (approximately 30 hr) to pre-train a Tacotron 2 model.

Speaker

A higher naturalness rating was anticipated for the adult speech synthesis due to the quality and quantity of the adult speech data used. Adult speech corpora of high quality are often more accessible than child speech corpora and the child speech data used in the current study was limited and recorded in a sub-optimal recording environment, which in part resulted in the child speech synthesis' decreased quality. Additionally, child speech differs quite substantially from adult speech, not only in terms of fundamental frequency, but also prosodic features. An adult's speech is often more fluent while a child's speech patterns are more inarticulate and vary significantly with regards to volume, pacing and emotional expressivity (Govender & de Wet, 2016; Jain et al., 2022). Although a MOS experiment is one of the traditional evaluation methods proposed by researchers (Govender & de Wet, 2016; Jain et al., 2022), the difference between adult and child speech may have also resulted in a lower naturalness rating for the child speech as naturalness is expressed differently between these speakers. It is fairly typical for a child to hesitate with new words, take breaks mid-sentence, mispronounce words and sounds, and wander off towards the end of sentences (Jain et al., 2022). It was hypothesized that listeners may have noted this distinctive child speech and prosodic characteristics and thought that the synthetic child voices were less natural as a result, likely due to different quality expectations for synthetic voices. Despite this, if the synthetic child voice on a speech-generating device is intelligible but still presents with minor mispronunciations or articulatory idiosyncrasies transferred from a donor child's speech, it may be more relatable for children who will be using the synthetic speech to communicate, as opposed to using adult speech. Research from Begnum et al. (2012) shows that children with expressive communication difficulties want voices that match their vocal identity, but not at the cost of intelligibility, as listeners must be able to understand them.

Similarly, researchers outlined that communication partners of children with expressive communication difficulties often prioritize intelligibility of the synthetic speech output, over similarity to the child's natural voice (Baxter et al., 2012; Begnum et al., 2012). Although the vocal identity of the child is still incredibly important to children's communication partners, it has been found that the synthetic child voice must be functional in various demanding environments, such as classrooms (Begnum et al., 2012). In addition, Drager et al., (2010) found that shorter sentences and single-word utterances produced with synthetic speech, are often less intelligible to listeners. Although our research has shown that listeners did not experience a significant difference in naturalness between sentences of differing lengths when modern speech synthesis was used, intelligibility may however still be affected, as this was not independently tested in the current study. Practically, using a single-word AAC device is very common for new AAC users, which suggests that communication partners may find it difficult to easily understand the synthesized speech output on entry-level speech-generating devices used by children unless the context is given (Drager et al., 2010; Terblanche et al., 2022).

Although synthesized SAE is likely new to listeners, listeners have probably heard synthetic adult US-English speech before via the internet and on popular social media sites. The SAE adult voice had a standard SAE dialect, which listeners may have been expecting, while the SAE child voice had a non-standard dialect. Despite the SAE child voice presenting with minor distortions and some noise, it is hypothesized that the English child voice was rated more harshly by listeners due to the child donor's unusual prosodic patterns and strong dialectal influence, which may have impaired the comprehensibility of the voice to unfamiliar listeners. On the other hand, Afrikaans and isiXhosa synthetic speech are not as freely available, so listeners likely did not have any predetermined expectations. As listeners aren't used to hearing synthetic speech in these languages, and even though dialectal variations were present, the child voices were not judged as harshly as the SAE. Results from the current study highlight that open-source software, such as Tacotron 2, opens the door for researchers to develop culturally, linguistically and age- appropriate voices in under-resourced languages and dialects, particularly for children with expressive communication difficulties.

Language

Despite the need for synthetic child voices, generating a synthetic child voice in under-resourced languages can be complicated, particularly when children's literacy skills are reduced. African home languages are mainly spoken, rather than written. Thus, children's literacy skills in African languages are often limited, due to the large language and literacy disparities in South African schools (Coetzee-Van Rooy, 2012). To account for this mismatch, spontaneous speech samples are often collected as an alternative. However, it is generally known that there is more variation, both acoustically and linguistically, with spontaneous speech as compared to read speech (Tucker & Mukai,

2023). In our study, although the isiXhosa child's speaking rate increased during the spontaneous speech sample, the sample was filled with partial words and repairs, resulting in less precise pronunciation. She also presented with a greater degree of code-switching during spontaneous speech than read speech. Ultimately, the data from the read speech was preferred for all three of the language models. Due to the challenges associated with collecting usable child speech data and the lack of applicable child speech corpora in different languages (Jain et al., 2022), it is not surprising that text-to-speech for child voices, particularly in under-resourced languages, are currently limited.

However, after the collection of child speech data and due to cross-lingual transfer learning, limited modifications are required for each language when using the default Tacotron 2 architecture. Transfer learning is an approach where a model trained on a source domain (i.e., English speech data) is used to improve the generalizability of a target domain (i.e., isiXhosa speech data) (Wang & Zheng, 2015). This is incredibly encouraging for the multilingual South African population. Non-English children with expressive communication difficulties often have to use English AAC devices (Tönsing et al., 2018). However, when there is a language mismatch between the language used on the AAC device at school and the language used at home, caregivers are less likely to use speech-generating devices (Tönsing et al., 2018). This language mismatch ultimately causes reduced buy-in from caregivers, modeling opportunities and consistent AAC use. Our study highlights that researchers will be able to either collect residual speech from children with expressive communication difficulties, with various home languages, and incorporate it into a training model, or if speech is severely impaired, they will be able to use an age-matched child's voice to develop synthetic child speech in under-resourced languages.

Warm Start Type

By implementing a warm start using available adult speech data, with Tacotron 2 (Wang et al., 2017), age, language, dialect, and gender mismatch are no longer limiting factors for the creation of suitable synthetic speech for children with expressive communication difficulties. For example, the final output of the SAE and Afrikaans child voices were male, while the training data used in the warm start was that of an adult female. When there are limited training data, which often occurs with child speech data, along with possible computational resource constraints, incorporating a warm start, with an established adult model of high quality, can improve the quality of the synthesized child speech output. The results from the current study support Phuong et al., (2021) findings, as the model's training time was reduced, and each model's quality improved dramatically after initializing a warm start.

It was anticipated, however, that including a warm start procedure twice, may result in the child model underperforming. For instance, the established Tacotron 2 model was used to warm start the

different South African adult models. The South African adult models (warm start B) were then utilized to warm start the respective child synthesis, in each language. Interestingly, the child models performed equally well with warm start B as with warm start A. This suggests that more than one warm start does not necessarily affect performance. Rather, the naturalness of the child model is directly proportional to the quality and the quantity of the adaptation data used. Although both warm start procedures produced adequate quality child speech synthesis, if one were to use the voices for AAC purposes, it would be sensible to choose the model that produces synthetic speech with the greatest consistency. As Tacotron 2 (Wang et al., 2017) is open-source and produces naturalistic synthetic speech, the lack of extensive speech data should no longer limit the development of appropriate child voices.

Child Speech Training Data

Shivakumar and Georgiou's (2020) study showed that any amount of child speech data were helpful, but 35 min of child speech data were able to give improvements of up to 9.1% over their adult model. Our results support Shivakumar and Georgiou's conclusions. Even though our study suggests that as little as 5 min of clean training data can produce a speech-like output when using a warm start procedure, the quality and accuracy of the synthetic speech systematically improve with over 1 hour of speech data. It is clear that the more high-quality speech data you have for training and adaptation, as seen with our adult South African voices, the more successful the model will be, which is supported by numerous researchers (Hasija et al., 2021; Kumar & Surendra, 2011; Shivakumar & Georgiou, 2020). Additionally, if the donor child's speech were to be recorded in a sound-attenuated room, one would get improved child speech synthesis results. Moreover, a scoping review by Terblanche et al. (2022) highlighted that the best synthesis results are usually found when data is carefully selected, rather than through blind selection. Similarly, in the present study, it was found that training Tacotron 2 with clean data (minimal disfluencies and errors etc.) provided a cleaner voice output. In other words, even if there were child speech corpora in under-resourced languages, researchers would need to spend time carefully selecting and cleaning the data, otherwise the articulatory errors, incomplete sentences, disfluencies, and hesitations that typically present in child speech would also appear in the synthetic speech output (Kumar & Surendra, 2011). Although cleaning the data has traditionally been considered a tedious task, our study shows that we no longer need over 24 hr of speech data to develop synthetic speech of adequate quality, due to the implementation of a warm start. Thus, reducing the amount of speech data needed also reduces the time spent cleaning the data. Our hope is that researchers will consider including a warm start method to develop synthetic child speech in other under-resourced languages.

Implications

Using this technology, and incorporating a warm start procedure, has the potential to support marginalized AAC communities and develop much-needed resources in low- and middle-income countries, especially where child speech data in under-resourced languages is restricted. This study showed that it is possible for children with expressive communication difficulties to have linguistically, culturally, and age-appropriate synthetic voices.

Limitations and Future Directions

It could be argued that using the established Tacotron 2 architecture is not in itself a novel approach to speech synthesis; however using modern text-to-speech technologies to generate text-to-speech resources in under-resourced languages for children with expressive communication difficulties, who live in low- and middle-income countries, has not received sufficient attention in the literature (Govender & de Wet, 2016; Hasija et al., 2021). We hope that this contribution can provide a model for others interested in creating synthetic child voices in under-resourced languages. Second, there are some limitations to the subjective quality assessments used. For instance, as the listening test was conducted online, some participants may have experienced technical difficulties, such as an increased lag time, which may have influenced their MOS rating. In addition, due to limited synthetic speech in under-resourced languages, fluent speakers have had few opportunities, if any, to hear synthetic speech in their home languages and dialects. Feedback from first language speakers is therefore particularly intriguing when considering future speech synthesis development in low- and middle-income countries. However, as only 124 listeners participated in the subjective MOS survey, with only 19 isiXhosa listeners, results are not necessarily generalizable to the entire population.

As an additional limitation, there are limited AAC devices/apps that are currently equipped to load the African synthetic voices we developed. Because the syntax and morphology of these languages differ greatly from English, researchers and clinicians will need to develop core vocabulary for each of these languages. Using a new or established device or app, this core vocabulary can then be paired with the synthetic speech in each language, to form a functioning speech-generating system. Along with the core vocabulary, one would need to incorporate graphic symbols with cross-cultural readability into a system to support children who are illiterate, or one could simply utilize a simple text-to-speech system in each language. However, using a system without symbols may limit its efficacy for many of the South African children with special needs, due to the low literacy levels amongst this population. Future research could incorporate the speech of children with expressive communication difficulties into the training procedure, so that synthetic voices can be even more individualized for specific children (Mills et al., 2014). In addition, obtaining feedback from children who make use of speech-generating devices

about their preferred synthetic voice (i.e., an adult, a child, or a child voice that includes some of their own speech) would be an important next step. It may be worthwhile for researchers to explore code-switching in speech synthesis, and to include more data to improve the pronunciation of names. If a high-quality child model is available, it may also be interesting to pre-train a model with child data rather than adult data. Lastly, it would be beneficial to gather in-depth feedback about the quality and intelligibility of the synthetic voices from caregivers, teachers, and speech and language therapists who have or work with children who have expressive communication difficulties.

Conclusion

This paper addressed the feasibility of creating synthetic child speech, using a limited training data scenario for three under-resourced South African languages, namely Afrikaans, isiXhosa and SAE. A small amount of child speech data was collected, and together with adult speech corpora, was used to pre-train a Tacotron 2 model. Despite limited child speech data, we were able to develop child voices that listeners rated as more natural than not. In addition, although the synthetic adult voices developed in this study appear more natural than the synthetic child voices, likely due to the quality and quantity of the adult data used, the child voices do match the voices of the South African donor children. In conclusion, this study highlights that due to recent technological advances, it is possible to accurately develop synthetic child speech with limited training data. This means that it is feasible to develop synthetic child voices in under-resourced languages for children with expressive communication difficulties, living in low- and middle-income countries, who may wish to speak in a language other than American or British English.

Supplemental Material

Table 1

Overview of the Training Data Utilized in the Development of Tacotron 2 Models for South African Adult Speech Synthesis in Three South African Languages

Language	Adult data (clean)	Utterances in training data	Warm start materials
Afrikaans	06:45 hr	3,868	Established Tacotron 2 model: LJ Dataset (adult North American English)
isiXhosa	06:44 hr	2,561	Established Tacotron 2 model: LJ Dataset (adult North American English)
South African English	07:02 hr	4,050	Established Tacotron 2 model: LJ Dataset (adult North American English)

Note. The table includes details on the language, duration of clean adult data collected, number of utterances in the training data, and the warm start materials used for model initialization.

Table 2

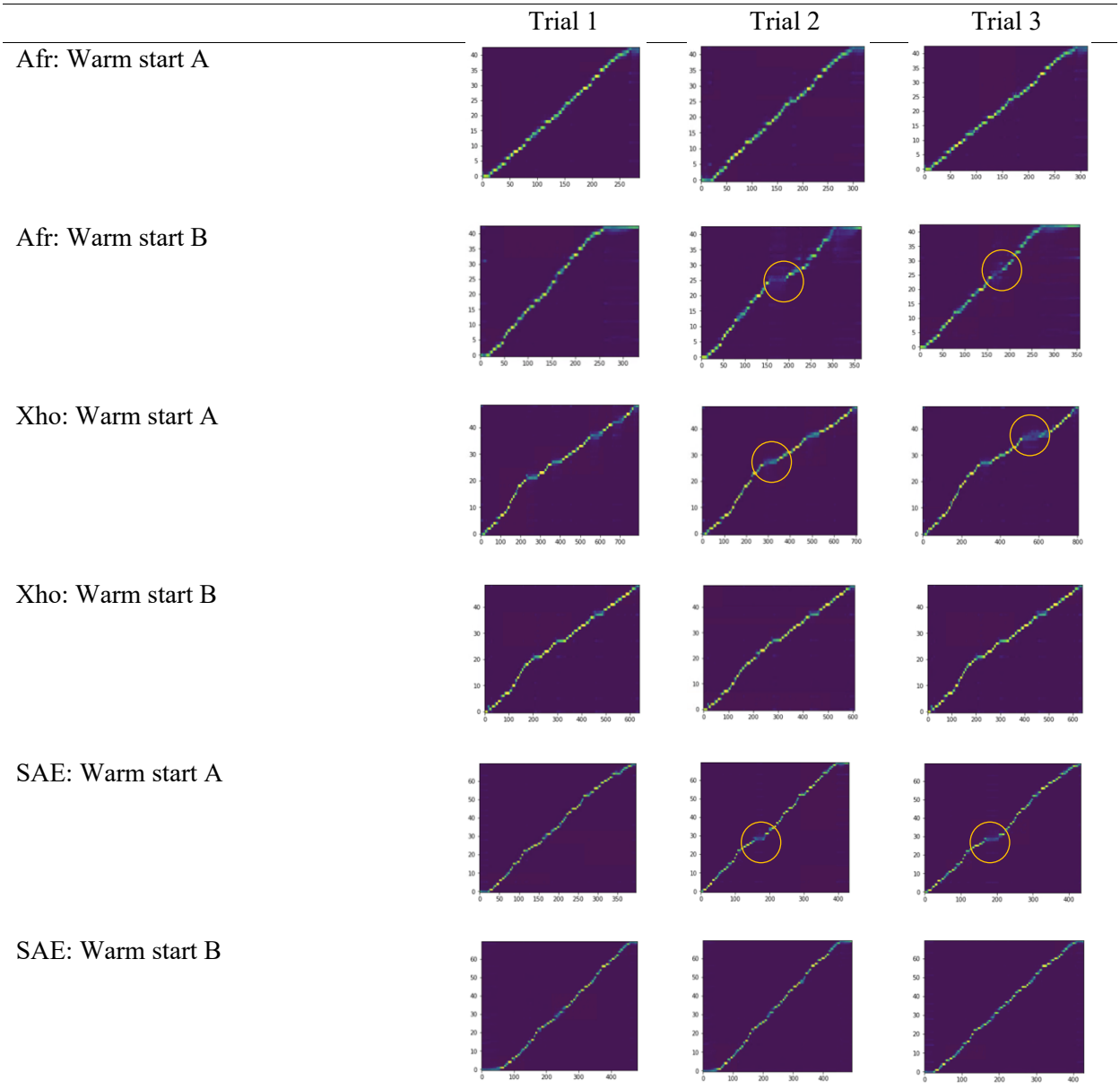
Overview of the Training Data Utilized in the Development of Tacotron 2 Models for South African Child Speech Synthesis in Three South African Languages

Language	Child donor	Child data (clean)	Utterances in training data	Warm start A materials	Warm start B materials
Afrikaans	One 12-year-old male	108.07 min	1,945	Established Tacotron 2 model: LJ Dataset (adult North American English)	Pre-made Afrikaans adult model: Lwazi III TTS dataset – Afrikaans
isiXhosa	One 12-year-old female	101.90 min	1,434	Established Tacotron 2 model: LJ Dataset (adult North American English)	Pre-made isiXhosa adult model: Lwazi III TTS dataset – isiXhosa
South African English	One 11-year-old male	113.70 min	1,949	Established Tacotron 2 model: LJ Dataset (adult North American English)	Pre-made SAE adult model: Lwazi III TTS dataset – English

Note. The table includes information on the language, characteristics of the child donor, duration of clean child speech data collected, number of utterances in the collected child speech data used for training, and the two warm start materials used for model initialization.

Figure 1

Alignment Plots Highlighting the Consistency of the Child Speech Models when Different Warm Start Procedures are Implemented (inconsistency circled)



Note. Figure 1 shows the alignment plots between the synthesized child speech models in each language, on three different occasions. I.e., the TTS in trial one was synthesized in the morning, the

TTS in trial two was re-synthesized at midday and the TTS in trial three was re-synthesized in the evening. Each point on the alignment plot represents a pair of corresponding elements, i.e., the synthesized speech and the donor speech. The blurred lines (highlighted sections) show where the model does not align well with the donor speech with regards to intonation, rhythm, and specific phonetic features, and is therefore considered inconsistent.

MANUSCRIPT FOUR

Do you like my voice? Stakeholder perspectives about the acceptability of synthetic child voices in three South African languages.

Camryn Terblanche, Michelle Pascoe, and Michal Harty

Division of Communication Sciences and Disorders, University of Cape Town, SA

trbcam001@myuct.ac.za/ michelle.pascoe@uct.ac.za /michal.harty@uct.ac.za

Abstract

Background

There is a global need for synthetic speech development in multiple languages and dialects, as many children, who cannot communicate using their natural voice, struggle to find synthetic voices on high-technology devices that match their age, social and linguistic background.

Aims

This study aims to document multiple stakeholders' perspectives surrounding the quality, acceptability, and utility of newly created synthetic speech in three under-resourced South African languages, namely South African English, Afrikaans, and isiXhosa.

Methods & Procedures

A mixed methods research design was selected. After the creation of naturalistic synthetic child speech which matched the vocal identity of three children with expressive communication difficulties, those three children answered questions about the quality, acceptability and utility of the synthetic voices using a pictographic 3-point scale. Eleven adults who are known to the children, participated in subjective quality assessments in the form of mean opinion scores, intelligibility tests and focus group discussions.

Outcomes & Results

Despite the synthetic adult voices appearing more natural, stakeholders were accepting of all the synthetic voices. Although personalisation of the voices is important, intelligibility is prioritised, and

standard dialects are often preferred. When communication partners have adequate training, are willing to model and support children in all environments, children with expressive communication difficulties thrive, but when AAC use is inconsistent, there is reduced vocabulary development and poor system transitioning, AAC abandonment is greater.

Conclusions & Implications

This research suggests that stakeholders from low- and middle-income countries are interested in the development of synthetic voices in their home languages. Our research highlights that children would prefer to incorporate these voices on their high-tech devices, and adults would prefer them for their children, learners, and/or clients' devices, rather than using British or US English voices.

What this paper adds

What is already known on this subject

- Caregivers, service providers, peers, and other communication partners play a substantial role in a child AAC user's early communicative success, and their acceptance of AAC ultimately influences the effectiveness of intervention. When communication partners advocate and support the inclusion of specific speech-generating devices, AAC applications, and suitable synthetic voices, children are more willing to consistently utilise the technology.

What this study adds

- As literature focusing on stakeholder perspectives from low- and middle-income countries is less common than research from high-income countries, and often lacks input from multi-perspective stakeholders, our study offers a unique perspective from South African children with expressive communication difficulties, caregivers of those children, their speech-language pathologists, and teachers, about the quality, acceptability, and utility of synthetic speech in under-resourced languages.

Clinical implications of this study

- Our research highlights that stakeholders would prefer South African languages and dialects on South African speech-generating devices, rather than relying on devices that only incorporate British or US English voices. The development of synthetic speech in under-resourced

languages has the potential to support marginalized AAC communities. Children with expressive communication difficulties would finally be able to participate in class and do so with a voice that matches their age, gender, social and linguistic background. This paper highlights the importance of providing a variety of synthetic voice options and emphasises the significance of introducing novel voices for high-tech AAC to children in a manner that respects and aligns with their linguistic and cultural backgrounds.

Introduction

There is a global need for synthetic speech development in multiple languages and dialects, as many children, who cannot communicate using their natural voice, struggle to find synthetic voices on high-technology devices that match their age, social and linguistic background (Mills et al., 2014). Language is a core component of individual and collective identity, and for children with expressive communication difficulties, the ability to communicate in a voice that resonates with their cultural background is crucial (Ndlangamandla, 2010). Children with expressive communication difficulties vary widely, each facing a unique set of language, speech, and communication challenges. Children may present with neurological disorders such as cerebral palsy, intellectual disability, autism spectrum disorder, genetic conditions like Down syndrome, and structural abnormalities such as cleft palate (Boesch & Da Fonte, 2019). Augmentative and alternative communication (AAC) provides vital support for these children. AAC serves as a tool to supplement existing speech or as an alternative method when a person is unable to use their natural speech to meet their communication needs across a variety of contexts and communication partners (Dada et al., 2022). Furthermore, learners with special educational needs (LSEN), including those with various challenges beyond expressive communication difficulties, can benefit from AAC. AAC helps them express themselves better, participate more in school activities, and interact socially (Boesch & Da Fonte, 2019). Despite improvements in medical treatments and technology over the last 30 years, the number of children under the age of 5 with developmental disabilities remains high. In 2016, approximately 52.9 million children globally (95% uncertainty interval [UI]: 48.7–57.3; or 8.4% [7.7–9.1%]) had developmental disabilities, with 54% of them being males. This is only a slight decrease from 1990, when an estimated 53.0 million children (49.0–57.1; 8.9% [8.2–9.5%]) had such disabilities. Notably, about 95% of these children resided in low- and middle-income countries (Global Research on Developmental Disabilities Collaborators, 2018). In South Africa, the prevalence of communication disabilities, which could benefit from AAC, is estimated to be between 6% and 12% amongst children aged over 5 years in South Africa (Dada et al., 2022).

Considering the high remaining prevalence, it has become even more important to find suitable AAC options for children. Although the range of high-tech AAC devices and applications are rapidly developing, without significant support from the children's communication partners, child AAC users are often unable to exploit the potential that these technologies can offer (Kent-Walsh et al., 2015; Moorcroft et al., 2019). Children who use speech-generating devices generally rely on applications that convert text or pictographic symbols into speech, using an impersonal standardised synthetic male or female voice (Nathanson, 2017). Although some devices allow for some personalisation, with regards to age or dialect, most easily accessible commercial speech technology often fail to recreate the unique variability of individual speech. Cultural variation within languages, such as dialectal differences or

regional accents, also plays a significant role in shaping how children perceive their own voices. When AAC users are unable to select a synthetic voice that aligns with their linguistic and cultural background, it may result in feelings of isolation, as they are unable to relate to the voice on their device (Mills et al., 2014). Synthetic child speech options in under-resourced languages, such as those spoken in low- and middle-income countries, are often limited or of reduced quality. Therefore, synthetic adult voices in major languages, such as English, are often selected for children who use speech-generating devices in these contexts (Tönsing et al., 2019). In South Africa, there are many official, as well as unofficial languages and dialects, with a considerable user base, that are not represented on South African speech-generating devices (Sefara et al., 2019; Tönsing et al., 2018). There is therefore a need to leverage the recent technological advances to improve the uptake of multilingual synthesis innovations.

However, if children do not relate to the voices used on their devices, it may lead to low technology adoption rates and withdrawal from social interaction (Mills et al., 2014), causing feelings of social isolation and loneliness (Nathanson, 2017). Enabling individuals with expressive communication difficulties to use synthetic speech in their home languages, with a voice that matches their vocal identity, can therefore have a profound impact on their quality of life. However, due to the distinct acoustic and linguistic properties of children's speech, finding suitable child speech corpora, appropriate for the development of synthetic child speech, can be difficult. Moreover, as researchers do not have access to the same level of text and audio resources in under-resourced languages, creating synthetic voices that appear natural and intelligible is even more challenging when under-resourced languages are used (de Wet et al., 2017) for synthetic child speech development. Due to these challenges, determining the acceptability of newly created synthetic child speech is an essential part of the development process.

Baxter et al. (2012) conducted a systematic review of the barriers and facilitators to the provision and use of high-tech AAC systems in various countries. They found several reasons why high-tech AAC might be rejected, for instance, decreased ease of use and reliability, inconsistent communication partner responses, poor perceptions, and a lack of AAC family and staff training, as well as limited synthetic voice and language options, especially when the language used on the device does not match the language used at home (Baxter et al., 2012), which is often the case in low-and middle income countries. However, contrary to Western countries, there is less regard for autonomy in Africa, as rights are centred around communal good and maintaining the continuity of relationships and interdependence shared within a community (Ganya et al., 2016). This implies that decisions concerning the child, including their ability to use a speech-generating device with a particular language, are discussed, and regularly determined by the community to which the child belongs. The

community's acceptance of a speech-generating device, and the synthetic voice therein, is therefore vital.

When communication partners advocate and support the inclusion of specific speech-generating devices, AAC applications, and suitable synthetic voices, children are more willing to consistently utilise the technology (Light et al., 2019). Intervention effectiveness ultimately relies on acceptance and modelling of AAC by communication partners (Moorcroft et al., 2019) and they therefore play a substantial role in a child AAC user's early communicative success. As the AAC landscape in low- and middle-income countries is so different to that of high-income countries (Mukhopadhyay & Nwaogu, 2009; Tönsing et al., 2019; Tönsing & Dada, 2016), and considering the role that stakeholders play in the acceptance of AAC (Moorcroft et al., 2019), it is important to gather stakeholder views and perspectives surrounding synthetic voices created for children with expressive communication difficulties in these contexts. This study therefore aims to document multiple stakeholders' perspectives surrounding the quality, acceptability, and utility of newly created synthetic speech in three under-resourced South African languages, namely South African English, Afrikaans, and isiXhosa. Stakeholders from low- and middle-income countries are often underrepresented in research, making our study distinct from those originating in high-income countries.

Method & Procedures

Research design

A mixed methods research design was selected for this study. Specifically, a triangulation mixed method design was selected as it allowed the researcher to expand on the quantitative results with qualitative data (Creswell & Clark, 2007).

In a study by Terblanche et al., (in press), naturalistic synthetic child speech was created. We collected 2 hours of child speech data per child from three typically developing children, who matched the vocal identity of the three children with expressive communication difficulties, including their age, gender, language and dialect. Using open-source adult speech corpora, we developed synthetic adult speech models in three languages. By combining these adult models with the child speech data, we created naturalistic synthetic child speech for each language using Tacotron 2 (Wang et al., 2017). The three children with expressive communication difficulties, for whom the synthetic speech was created for, answered questions about the quality, acceptability and utility of their newly created synthetic voices using a pictographic 3-point scale in this manuscript. Eleven adults who were known to the three children, participated in subjective quality assessments in the form of mean opinion scores (MOS), intelligibility tests and focus group discussions. One focus group included caregivers of children with

expressive communication difficulties whilst another included teachers and speech-language pathologists (SLPs) who work with children with expressive communication difficulties. Two focus groups were conducted to ensure that power imbalances did not occur between participants.

Following the guidelines provided by the Helsinki Declaration of 2013 (World Medical Association, 2013), this study obtained ethical approval from the Human Research Ethics committee (HREC no. 765/2021) at a South African university. All participants were asked to maintain confidentiality, and each participant had to be legally competent and informed before they were invited voluntarily to sign a consent form.

Participants

Purposive sampling methods were used. Researchers intentionally selected participants who have expressive communication difficulties, as well as those who work, consult with, or have children with expressive communication difficulties. To reflect South Africa's rich multilingual culture, participants with different linguistic backgrounds were selected. Most of the participants were recruited from three LSEN schools in Cape Town, while two of the professionals were recruited via social media. Once the three children were selected, the adult participants were selected by moving outwards in concentric circles. The adult participants were therefore familiar with at least one of the child participants, including their natural voice, social and linguistic background, as well as their specific needs and difficulties.

Children

Three children with expressive communication difficulties, for whom the synthetic speech was created for, participated in the study. The children's diagnoses include autism spectrum disorder (Afrikaans male, 11 years old), significant intellectual disability (English male, 11 years old), and cerebral palsy (isiXhosa female, 12 years old).

Caregivers

Eight caregivers consented to participate, but only four were able to attend on the day. The mean age of the caregivers was 47 years old. Two participants were mothers of children with expressive communication difficulties, one was an aunt, and another was a grandmother of a child with expressive communication difficulties.

Professionals

Eight teachers and SLPs consented to participate, but only six participants were available to participate on the day. The mean age of the participants in the professional focus group was 38 years old. Table 1 outlines the participants' key sociodemographic characteristics. Two were teachers in public LSEN schools; three were SLPs working in public LSEN schools, and one was an SLP working in private practice, but with experience in LSEN contexts.

Table 1

Professional Focus Group Participant Demographics (n=6).

Demographics	Number of professionals
Age range	
20-30	2
31-40	2
41-50	1
51-60	1
Sex	
Male	1
Female	5
Languages used in intervention/classroom	
English	1
Afrikaans/English	2
Afrikaans/IsiXhosa/English	3
Years of experience	
3-5	1
6-10	2
11+	3
Service sector employment	
Public	5
Private	1
Employment type	
Teacher	2

Data collection

The children's perceptions of the synthetic speech were each collected separately at their respective schools. After listening to the synthetic speech relevant to their home language, each child answered simple questions using a pictographic 3-point scale (i.e., yes, not sure, no) (see supplementary material). On the other hand, data from each focus group session (professionals vs. caregivers) were collected at a neutral location after working hours, and each session lasted approximately two hours each. Each focus group session was divided into three sections.

Mean opinion score

For the adult participants, the first section involved participants listening to audio clips and giving their mean opinion scores (MOS). The MOS method is regularly used in speech synthesis as a valid method to subjectively evaluate the performance and quality of synthetic voices, as observed in Govender and de Wet (2016), Jain et al., (2022), and Sefara et al., (2019). Participants listened to the synthetic speech from one device with a loudspeaker, rather than through headphones in a controlled environment, as this scenario matches the functional application of the voices in the real world. For each language, participants listened to the synthetic child speech before listening to the synthetic adult speech. Participants were asked to give a rating using a 5-point Likert Scale ranging from 1 (*horrible/completely unnatural etc.)* to 5 (*excellent/completely natural etc.),* which reflected their a) overall impression (quality), b) the pleasantness, c) naturalness, and d) understandability (degree of listening effort), as well as its e) similarity to real speakers. Participants had the flexibility to listen to the voices as frequently as they liked.

We chose to assess multiple factors because they collectively evaluate different yet related aspects of synthetic speech quality. Each factor provides unique insights crucial for understanding how well synthetic voices meet user expectations and perform in real-world applications. For example, pleasantness evaluates participants' subjective enjoyment and emotional response to synthetic speech, influencing its acceptance, while overall impression assesses the system's clarity and effectiveness in conveying information. Naturalness evaluates how closely synthetic speech mimics human-like qualities such as rhythm, stress and intonation. In contrast, similarity to real speakers evaluates how well the synthetic speech replicates the specific characteristics of the original donor's voice, which for the synthetic child voices, was chosen to match the vocal identity of the specific children with expressive communication difficulties, ensuring accuracy in replicating individual speaker nuances.

Intelligibility

Following this, if participants were fully proficient in the language, they were asked to participate in intelligibility tests and transcribe what they heard. Each language had six semantically predictable sentences, divided equally between synthetic child and adult speech. Research on child speech synthesis has shown that children's speech is difficult to interpret using semantically unpredictable sentences, which is generally used during conventional evaluation (Govender & de Wet, 2016). Therefore, semantically predictable sentences were used during evaluations. While there might be expected correlations between measures like WER for intelligibility and MOS for understandability, both were essential for a comprehensive assessment. WER provided objective accuracy metrics, indicating how accurately the speech was perceived in terms of its correctness or errors in reproduction. In contrast, MOS ratings captured subjective perceptions of ease and clarity in comprehension, reflecting practical usability in various communication scenarios, such as school environments.

Focus group discussion

After a short break, the third section included the focus group discussion. As all participants were comfortable conducting the discussions in English, a translator was not required. Different interview schedules were created for each group (professionals vs. caregivers). So as to guide the discussion, the facilitator asked topic-related descriptive questions, starting with general questions (e.g., "What do you think about the synthetic voices you just heard?") and proceeded to specific questions (e.g., "How do you think your family/the school community will view and interact with the child if they use these voices?") To ensure credible data, focus groups were audio recorded and immediately following each group, typed transcripts were prepared (Liamputtong, 2011). To maintain anonymity, no identifying information was recorded in the transcripts. Transcripts were provided to the participants for member checking. None of the participants indicated they required changes to the transcripts.

Data analysis

Two distinct focus groups were conducted with the adult participants, the data from both were analysed together and the children's data were also included. Firstly, each adult participant listened to and evaluated 14 unique audio clips per language. This included two synthetic child audio clips and two synthetic adult audio clips for MOS ratings (overall impression, pleasantness, naturalness, and understandability), two audio clips for assessing similarity to real speakers (comparing to donor recordings), and three audio clips each for evaluating the intelligibility of synthetic child and adult speech. This approach aimed to ensure that participants rated various dimensions of synthetic speech, whilst managing participant fatigue and ensuring focused evaluation. In total, each adult participant made 30 MOS ratings (10 MOS ratings per language) therefore, 330 MOS ratings were compared

overall. Each child participant answered 9 closed-ended questions to gather insights into their subjective experiences and the practical usability of the synthetic voices in their daily lives. All hard copy ratings were transferred to a .csv file. Secondly, using the adult participants' transcribed sentences and based on the formula provided in the Blizzard 2007 challenge guidelines (Clark et al., 2007), the participants' average word error rate (WER) was calculated for each model. Spelling mistakes and typographical errors were corrected before the WER was calculated. Using R (R Core Team, 2019), both descriptive and inferential statistics were run on the MOS and intelligibility data. To test the statistically significant effects, a linear mixed-effects model was implemented with the ordinal package (Christensen, 2022) in R (R Core Team, 2019).

Thirdly, verbatim transcripts of the focus group discussions were analysed via reflexive thematic analysis, as originally outlined in 2006 (Braun & Clarke, 2006, 2019, 2021), facilitated by NVivo (QSR International, 1998). Reflexive thematic analysis is a theoretically flexible interpretive method for qualitative data analysis, enabling the identification and examination of patterns or themes within a dataset (Braun & Clarke, 2019). The focus group sessions specifically aimed at discussing predefined evaluation factors for assessing synthetic speech quality, which were derived from previous researchers (Govender & de Wet, 2016; Jain et al., 2022; Sefara et al., 2019). This approach reflects a deductive thematic analysis, where predefined themes informed the analysis process (Braun & Clarke, 2019, 2021), ensuring systematic exploration and interpretation of participant insights. However, we remained open to new themes that emerged from the data itself, which ensured a comprehensive exploration of the research question (Braun & Clarke, 2019). This combination of deductive and inductive elements provided a balanced framework for understanding both anticipated and novel aspects of synthetic speech quality. Two judges independently applied the coding framework to the focus group data. Discrepancies were discussed until consensus was reached and 100% agreement between the two judges occurred, ensuring validity and reliability of the analysis and the identification of key information.

Results

Themes and subthemes that emerged from the analysis are presented in Table 2. All the information is presented as one dataset, despite conducting two separate focus groups, and collecting the children's responses separately.

Table 2

Identified Themes, Subthemes, and Key Messages from Participants

Themes	Subthemes	Instances identified	Key messages from participants
Quality of the voices	Naturalness and intelligibility	25	The English child voice is significantly less natural than the other synthetic voices.
	Rate of speech	7	A slower rate of speech might be more accommodating to the listener.
Personalisation of the voices	Dialect	15	Non-standard dialects might be disliked by families who do not have the same dialect.
	Age and identity	10	Children relate with the synthetic voices when they have the same age, gender, social and linguistic background.
Implementation and use	Barriers	17	One of the biggest barriers to high-tech AAC use is a lack of communication partner involvement.
	Facilitators	17	Low-tech AAC facilitates the transition to high-tech devices.

Three themes emerged from the data, namely, quality of the voices, personalisation of the voices, and implementation and use. The implementation and use of AAC, which included the subthemes, barriers and facilitators to the implementation and use of AAC, was most discussed by stakeholders.

Quality of the voices

The quality of the voices theme considers the speech synthesis' standard and suitability. Two subthemes were identified, 1) naturalness and intelligibility, as well as 2) the rate of speech. In addition, the children's perspectives and the adult participants' mean opinion scores are also discussed.

Each one of the children indicated that they liked the child voices created for them. The adult participants shared that they were pleasantly surprised by the naturalness and realistic prosodic features of the synthetic voices, "the [commercially available voices] are very broken up, whereas this was lovely and flowing". One SLP said:

> I had to remind myself that we weren't listening to real people. It didn't sound 'Stephen Hawking'. It sounded like you were listening to someone on Zoom. It was way more natural. If I didn't know it was a computer voice, I probably wouldn't have thought twice that it wasn't a [real] child.

Through the discussion, it became clear that adult participants preferred the Afrikaans and isiXhosa child voices. One SLP shared that they *really liked the isiXhosa voice* and thought it was the *most natural* while another said the Afrikaans voices were more natural compared to what is available, noting that they were *aeons ahead*. A caregiver stated that she really liked the Afrikaans child voice and described it as *very cute*. The adult participants' subjective evaluation of the quality of the voices validates their mean opinion score (MOS). The adult participants perceive the English child voice to be the worst quality ($\bar{x}$ = 3.1) and the least natural ($\bar{x}$ = 3.0). Figure 1 illustrates the average MOS when considering the adult participants' ratings for the overall impression ($\bar{x}$ = 4.05), pleasantness ($\bar{x}$ = 3.98), naturalness ($\bar{x}$ = 3.87), and similarity to real speakers ($\bar{x}$ = 3.83) for all the voices.

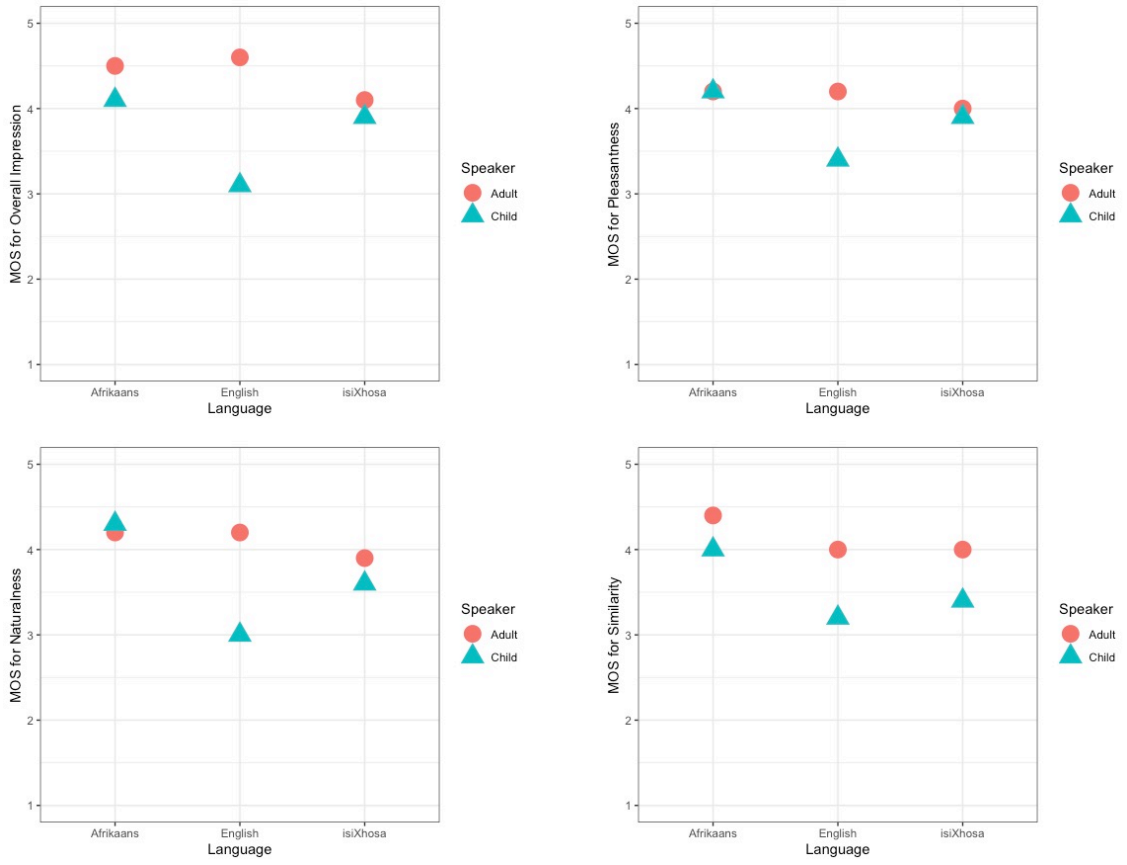


Figure 1

Mean Opinion Scores of the Overall Impression, Pleasantness, Naturalness, and Similarity to Real Speakers, with Reference to Speaker and Language

From Figure 1, there is a notable difference between the English adult and child voices with regards to the overall impression and naturalness of the voices. To test the statistically significant effects, a linear mixed-effects model was implemented with the ordinal package (Christensen, 2022) in R (R Core Team, 2019). When predicting the overall impression with the language and speaker, the model's intercept, corresponding to Afrikaans adult, is at 4.50. The effect of the English child speaker is statistically significant and negative (beta = -1.10, 95% CI [-1.92, -0.28], t(52) = -2.69, p = 0.009; Std. beta = -1.39, 95% CI [-2.43, -0.36]). When predicting the naturalness with the language and speaker, the model's intercept, corresponding to Afrikaans adult, is at 4.20. The effect of the English child speaker is statistically significant and negative (beta = -1.30, 95% CI [-2.10, -0.50], t(52) = -3.28, p = 0.002; Std. beta = -1.64, 95% CI [-2.65, -0.64]). There is not a statistically significant difference with regards to similarity and pleasantness. Figure 1 highlights that all the adult voices are usually considered superior to the child voices. However, the Afrikaans adult voice was considered slightly less pleasant than the Afrikaans child voice, but both are rated highly in each category. Very little difference is found between the isiXhosa adult and child voices, as both are rated highly in each category. Similarly to the adult participants, each child participant indicated that the synthesis pronunciations, matched their expectations for each respective language.

Although the understandability results from the MOS indicate that on average, minimum effort was needed to understand the child and adult voices in all languages ($\bar{x}$ = 4.2), one caregiver shared that in comparison to the child voices, understanding the adult voices just *came naturally*. The professionals agreed that the adult voices were *easy to understand*. Alternatively, the children were pleased with both the child and adult voices and indicated that other people would be able to understand them, no matter which voice (child/adult) they selected for their device. One SLP thought that the English child was *the hardest to understand*. A caregiver agreed and mentioned that the English child *wasn't so clear*. One SLP said, “it would make me think twice about using [the English child voice] because I think there are other voices that are clearer, but it totally depends on your kid.” A caregiver shared a similar sentiment and noted that if she had to pick one for her child, she would *pick an adult one.* They suggested that the voices should immediately be understood by communication partners, because “nobody goes back and [listens to] the thing twice. You just want to hear it once and know exactly what they are saying without putting in any effort”.

After the MOS analysis, it appears that all the adult participants agreed and believed the English child voice required remarkably more listening effort. To test the statistically significant effects, a linear mixed-effects model was implemented. When predicting understandability with the language and speaker, the model's intercept, corresponding to Afrikaans adult, is at 4.30. The effect of the English child is statistically significant and negative (beta = -1.20, 95% CI [-2.04, -0.36], t(52) = -2.87, p =

0.006; Std. beta = -1.45, 95% CI [-2.46, -0.44]). This statistically significant effect is illustrated in Figure 2 (overall understandability $\bar{x}$ = 4.16).

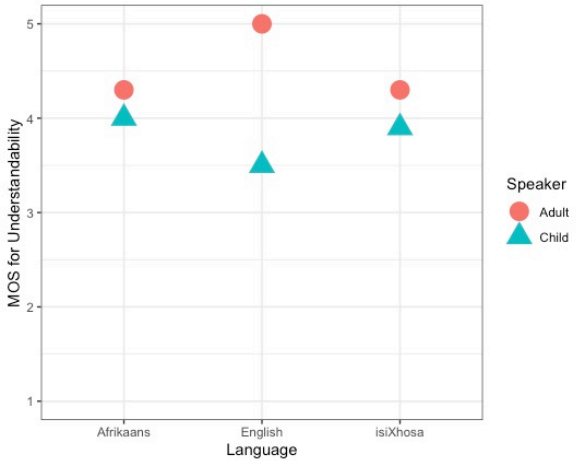


Figure 2

Understandability Mean Opinion Scores of the Synthetic Voices, with Reference to Speaker and Language

Adult participants also took part in an intelligibility test. The results were quantified in terms of the word error rate (WER) that occurred in every transcribed sentence (with typographical errors and spelling errors corrected). The average WER for all the voices was 4.93%, the average WER for the adult voices was 1.99% and the average WER for the child voices was 7.88%. The WERs corresponding to the different voices, are illustrated in Figure 3.

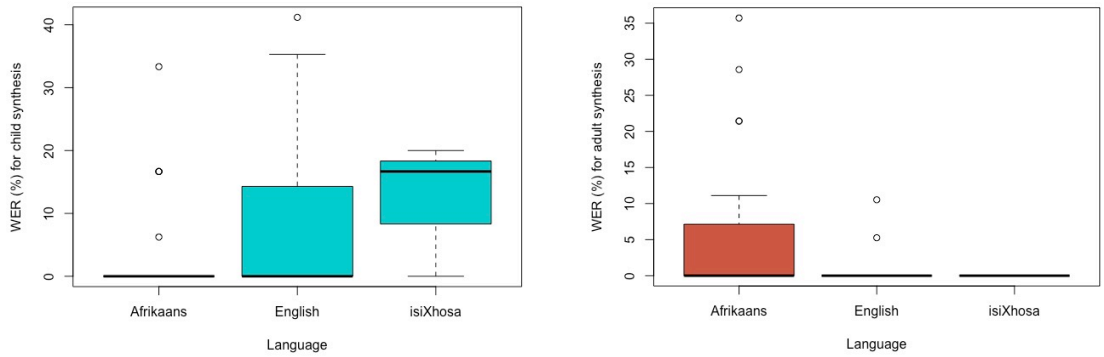


Figure 3

Boxplots showing the Participants' Mean Word Error Rate for the Synthetic Voices, with Reference to Speaker and Language

Similar to the understandability results illustrated in Figure 2, Figure 3 shows that the isiXhosa adult voice performed the best (WER= .00%), followed by the English adult voice (WER= 0.50%), with a negligible difference between them. Alternatively, the WER results show that the Afrikaans adult voice (WER= 5.48%) was the least intelligible adult voice and the Afrikaans child voice was the most intelligible child voice (WER= 3.2%), while the English child voice (WER= 8.43%), and the isiXhosa child voice (12. 00%) trailed behind. To test the statistically significant effects, a linear mixed-effects model was implemented. When predicting WER with language and speaker, the model's intercept, corresponding to Afrikaans adult, is at 5.76. The effect of the English child speaker is statistically significant and positive (beta = 10.19, 95% CI [4.03, 16.35], t(112) = 3.28, p = 0.001; Std. beta = 1.12, 95% CI [0.44, 1.80]) and the effect of the isiXhosa child speaker is statistically significant and positive (beta = 14.26, 95% CI [0.13, 28.39], t(112) = 2.00, p = 0.048; Std. beta = 1.57, 95% CI [0.01, 3.12]).

The adult participants also considered the rate of the speech and how that affected intelligibility. One SLP believed the isiXhosa child's voice "just [felt] slow… I could figure out [where] each word began, where it ended". A teacher believed that a slower rate of speech could be *more accommodating to the listener*, and *benefit* the learners with special education needs, particularly their *receptive language*. Another SLP agreed and shared that the Afrikaans child was *quite fast sometimes*, which affected her understanding of the Afrikaans speech. The English speech was *much closer to normal speed*.

Personalisation of the voices

The personalisation of the voices theme encompasses participants' perspectives surrounding tailoring the synthetic voices, with two subthemes: 1) dialect, as well as 2) identity and age.

The adult participants discussed the role that dialect plays in listeners' perceptions of people. The synthetic adult voices had standard dialects, while the synthetic child voices had non-standard dialects. Speaking about the English child voice, who presented with a prominent Cape Flats English dialect, one SLP said that if she had a client from that area, she would use it, but expressed that *some of the parents would not be happy*. She suggested that it may be a *cultural issue* as they would probably choose a British or an Australian voice instead. Similarly, another SLP believes voices with non-standard dialects, such as the Afrikaans child voice, "only serve one population… if my child doesn't brei (speak with a uvular r) or doesn't come from a family that brei's, I'm not going to give them that kind of voice." Evidently, it depends on the child's social background as another SLP stated that they would *definitely use the Afrikaans child voice* for one of their clients, as it would *suit him better* than the voice he currently uses. When comparing the Afrikaans child and adult voices, one caregiver who came from a similar background as the child, could relate to the voice and said that she "heard the

difference between [the] coloured[1] and [the] English person speaking…I like that [Afrikaans child voice]."

The isiXhosa dialect was well-liked by the participants, especially because they hadn't heard it before. One SLP expressed that the language barrier has been a significant challenge in their context, as a *child speaking US English in a fully isiXhosa class is odd*. In the end, the SLPs agreed that children with expressive communication difficulties should "choose their voice. If they want to sound like that kid with an accent that isn't from their area, and they think it's cool- they don't have a lot of control over their communication- so go for it!" The teachers believe that due to the social background of the children in their classrooms, the child voices are *easy to relate to*, as you've *heard the accent* that's been used, compared to the commercial voices, which makes it *a very useful tool*, especially with their specific learners in their classrooms.

It became clear during the discussion that a person's voice reflects their identity and age. One caregiver shared that, children listeners might relate to the child voice because they *think it's their peer*. Although two out of the three children believed that the synthetic child voices did not sound exactly like them, they were happy to use the voices and were excited to hear synthetic speech in their language. The two younger male children both selected the child voices for their devices, while the 12-year-old female selected the adult female voice for her device. When discussing the children using the adult voices, one teacher said that,

> The adult voices had some sense of authority behind them. [Having an adult voice for a child] would be strange because it won't necessarily allow [the children] to be themselves. Even if you know that this is a child, [in] the back of your head, you're like, it's an adult [speaking].

A SLP agreed that it might be *perceived negatively* if the child *had an adult-type voice with that kind of authority coming out*, but ultimately it comes back to quality as the participants believe *the adult voices were clearer* which is preferable. The professionals also feel that the child voices seem to be more enthusiastic, similarly to how typical children speak as they have *lots of intonation*, which made them think it *would be pleasant* for children.

Some of the child voices would *be a better fit* for children, based on their age. The professionals felt that the English and Afrikaans child voices sounded like children that were *maybe around eight to ten* while the isiXhosa child voice was perceived to be slighter older at *about thirteen and above*. As the

[1] The term "coloured" in South Africa has a specific historical and social context. Unlike its pejorative connotations in the UK and US, "coloured" is a recognised racial classification that originated during the apartheid era and is still used today to describe a distinct cultural group in South Africa. People who identify as coloured often have a mixed racial heritage. The participant's use of the term reflects a shared cultural background and identity, which is crucial for understanding the connection they felt with the child.

typically developing child voice donors were 11- and 12- years old respectively, the professionals were very close, which suggests that the voices closely reflect the ages of the children.

As children with expressive communication difficulties would finally be able to speak and participate in class with a synthetic voice that matches their social and linguistic background, their gender and age, caregivers believe that these synthetic voices would *boost the child instead of making him feel different*. Another caregiver shared that "I would be proud if I had a [device] like that [because] it was made especially for [me]".

Implementation and use

This theme considers why children might abandon or continue using high-tech AAC devices. Two subthemes were identified, including 1) barriers and 2) facilitators to implementation and use.

Participants believed that one of the biggest barriers is reduced communication partner involvement. If there is a *lack of process, thought and provision of effective vocab* by SLPs, or if teachers aren't *hands on*, the child might abandon the system. Caregiver involvement is also crucial as one SLP shared that, "the iPad often comes in uncharged, and you get ten minutes out of it." Although SLPs are sensitive to teachers' demands in LSEN classes, because *every single child has extensive needs*, they believe it is challenging to get buy-in and implement a high-tech AAC system in class. Generally, professionals believe that *inconsistency of the iPad* can cause the child to abandon the system.

Moreover, limited vocabulary, as well as limited AAC and vocabulary transitions, when the child's language inevitably develops, is another barrier to high-tech AAC use. One caregiver said that "[the vocabulary] should be upgraded all the time, like they do with your phones". A SLP agreed and mentioned that she likes to *keep a system for at least five years* but that children need *access to vocabulary* and they should be *transitioned* to a new AAC device *whenever necessary*,

> Kids get stuck with the same system from ages six to whatever… we need to be aware, are they becoming more literate? If they're not hugely physically disabled, [they'll] probably type faster than [they'll] be able to find words on an iPad. And in some ways, [using text-to-speech] is more socially acceptable.

Bullying and subsequent embarrassment are other reasons why participants' feel children might abandon their devices. A teacher said that if the other leaners *start teasing*, they *may stop using it*. This was also spoken about by a caregiver who shared that other family members would be *very jealous* and say things like "why must [child's name] get that? Why not us?" iPads that do not have *guided access*

are often misused by other family members, which limits use for the children who need it to communicate. One SLP said that older children might stop using devices even if they are useful because of "changes in mood or status…maybe you thought the device was going to help you in primary school and now you're going to high school and you have to take it, how embarrassing." Lastly, they may stop using AAC if they get fatigued as "AAC is slow and hard. Even an efficient AAC user is slower than a verbal speaker, and it's effortful."

There are multiple factors which encourage AAC use, such as communication partners' motivation and acceptance of AAC and training. All the children indicated that using synthetic speech would be helpful at home and at school. One SLP shared that with AAC training, "Probably 80% of my work is with her facilitator, her OT, her swimming coach, her mother, her everyone. I work with everyone around the child, probably to a greater degree than the child themself".

Teacher training and handover is essential, as some of the teachers who have had previous learners using AAC, feel a bit *more comfortable*. They mentioned that there's always *a refresher needed*, but *it's not so scary*. It's also helpful when teachers can *create their own boards*. One teacher who has never had a child use a high-tech AAC device in their class agreed to try because they noted that LSEN *teachers never really know if the learner understands* the instruction up until the learner has to write the actual activity. Peers can also assist in the classroom, one SLP shared,

> We used to have a buddy system with tech, so that the teacher can carry on teaching and another child can assist if possible… I think it helps to familiarise people with devices and it also takes the fear away of 'if I touch it, I'm going to get into trouble.' [AAC] is just in the classroom.

With that in mind, exposure is vitally important. Encouraging peers to handle the devices has been found helpful, increasing opportunities for peer modelling, as one SLP expressed how she loves it when the other leaners talk to each other and explore the app. In order to facilitate ease of use, having children in the *same class on the same app* is useful, as children can learn together. However, SLPs discussed that this can be tricky to manage in some classes where children have autism spectrum disorder and experience impulsivity, as *they want to touch the whole time* but the teachers shared that *social stories* can help. Additionally, participants believe that teachers are sometimes hesitant to use AAC in the classroom, so first exposing the teachers to *technology in the classroom* is like *a doorway to AAC*. One teacher said that as long as you can "show how [AAC] is going to help teach and assess children, they will be on board. I can guarantee that."

Similarly, if the classroom is arranged for AAC use, it makes transitioning to the high-tech AAC device that much easier for children, for example, "we have a basic core page that is available hardcopy

in the classroom, like a zero size, and then obviously the kids have [the core page] on their devices." Additionally, if the whole class is *introduced to between two and four words to work on for the week*, the SLP's *can supplement* and work with the child on whatever the teacher has to work on. This gives the professionals a practical process to follow, which is helpful, and once the children are more *comfortable using the device*, or as soon as they notice their improvements, they often want to continue using the device.

SLPs have found that many children using iPads in class have "become more verbal and are able to express themselves without the need for AAC… they have been discharged from it being their voice, and it's kept in the class for language support".

Discussion

Our study included South African stakeholder perspectives on synthetic child speech. Recent literature focusing on stakeholder perspectives from low- and middle-income countries is less common, and often lacks input from multi-perspective stakeholders (Mukhopadhyay & Nwaogu, 2009; Tönsing et al., 2019; Tönsing & Dada, 2016). It is clear from the study that stakeholders were accepting of all the synthetic voices in the different languages. The children would be willing to incorporate these voices on their own high-tech devices, and the adult participants would be willing to have these voices on their children, learners, and/or client's devices. Participants considered the overall impression, pleasantness, naturalness, understandability, and similarity to real speakers, and even though the adult voices were considered superior to the children's voices, none of the synthetic child voices were rated negatively or below "neutral" on the Likert Scale, in any of these categories. One of the suspected contributing factors to the reduced perception of quality for the synthetic children's voices was the decreased rate of speech and the increased degree of expressivity in the donor children's speech, which made the synthetic child voices appear as though they had unusual prosody. Moreover, as intelligibility and naturalness is expressed differently between adults and children, participants likely thought that the occasional mid-sentence break, minor mispronunciations, and intermittent hesitations in the synthetic child speech output was attributed to the synthesis system, rather than a typical child's voice, even though average child speech is known to have these distinctive child speech patterns and prosodic characteristics (Jain et al., 2022).

When considering the intelligibility of the synthetic voices, the results show that on average, listeners could understand the synthetic adult speech 98% of the time, but only 92% of the time for the child speech. Similarly to findings from Begnum et al.'s (2012) study, participants and particularly the caregivers, appeared to value intelligibility over all else. However, they were excited to learn that synthetic speech development is possible in different South African languages, reflecting different ages

and genders. The predominant experience from the adult participants in this study was that the English child voice, which had the strongest dialectal influence, was considered the poorest quality and the least intelligible synthetic child voice. This was not the case with the perception of, for example, the Afrikaans and isiXhosa synthetic child voices which each had a less prominent albeit non-standard dialect. Although synthetic speech with non-standard dialects is less common due to the lack of speech data in these dialects (Terblanche et al., 2022), our study suggests that listeners may be more accepting of standard dialects (i.e., non-dialectal, non-accented speech), on speech-generating devices. This may be because a) participants experience improved comprehensibility with dialects and accents that they have greater exposure to (Perry et al., 2018), b) the standard dialect matches their previous assumptions of what synthetic voices should sound like (speech synthesis systems like Alexa and Siri produce standard speech), c) listeners perceive the standard varieties to be more prestigious or they may prefer the standard dialect as d) the non-standard dialect reflects a distinct racial, social, or cultural group that they are not part of (Gates & Ilbury, 2019). It has been found that due to the similarity attraction effect, people are more trusting and accepting of voices when it matches their own dialect or accent (Pucher et al., 2009).

Even though standard dialects appear to be preferred by the professionals and caregivers in our study, it is clear from the results that introducing synthetic speech with differing dialects offers children with expressive communication difficulties the power of choice, which is empowering. Some adult participants in our study were able to really relate to the synthetic speech with the non-standard dialects, while others could immediately determine which clients would benefit or refuse to have these voices on their systems. As dialect is an indicator of a child's unique background, children might feel a greater sense of belonging to their community if they are able to select their respective dialect. Language also plays a role in identity and belonging, particularly in a multilingual country such as South Africa (Tönsing et al., 2019). Instances where multiple children share the same voice, even if it differs from their home language, highlight the importance of ensuring that the synthesised voice aligns with the AAC user's identity (Mills et al., 2014). When children with expressive communication difficulties aren't given access to AAC resources, interventions, and systems in their home languages, they often cannot engage with the academic content at school (Tönsing et al., 2019), and they also lose a significant part of their cultural heritage (Ndlangamandla, 2010). Professionals in our study highlighted that due to the progressive nature of AAC learning, when a child has access to speech synthesis systems in their home language, both their receptive and expressive language is supported. However, given the scarcity of diverse voice options available, it's vital to recognise the challenges child AAC users may face in finding voices for their AAC devices that match their background or identity. This shortage could be particularly problematic in classrooms with multiple children needing AAC support.

Although it is more challenging to create synthetic child speech in comparison to adult speech, due to the lack of child speech data, the importance of having a synthetic voice that reflects what the child believes their natural voice to sound like, including their age and gender, cannot be overemphasised (Ripat et al., 2019). The results from the current study support this, as the children were able to play a role in the decision making. The oldest female child selected the adult female voice for her device and as she is currently transitioning into a teenager, she may have felt like an adult voice was more relatable. Article 12 from the Convention on the Rights of a Child (United Nations, 1989) asserts that therapists are duty bound to include children in the decisions that impact them, in accordance with a child's age and maturity. Children with expressive communication difficulties have the right to have their opinions heard when it comes to selecting a synthetic voice, and the ensuing choice of their language, gender, age, and dialect. While caregivers and therapists should clearly also be closely involved in decision making for a child, a child's opinions should hold equal weight.

Communication partners, such as caregivers, teachers, SLPs and peers play a fundamental role in a child's willingness to accept an AAC system (Kent-Walsh et al., 2015; Moorcroft et al., 2019). Participants in our study shared that if communication partners have adequate AAC training and greater exposure, then children with expressive communication difficulties thrive. Conversely, participants believe that inconsistency of AAC use, reduced vocabulary development and poor system transitioning were barriers to high-tech AAC use. Additionally, our study highlighted that children with expressive communication difficulties occasionally fall victim to bullying and teasing, which may cause AAC abandonment. The professionals in our study believe greater peer exposure to AAC may reduce teasing, and it also supports all the children by setting up classrooms for AAC use. Peers are not only able to successfully model AAC use to children with expressive communication difficulties, but they are also able to assist when troubleshooting is required, which ultimately improves their own language and digital literacy skills (Finke et al., 2009). Professionals shared that having a low-tech AAC system in the classroom may encourage early attempts at accessing the system, which should make transitioning to high-tech AAC systems that much simpler. Lastly, the professionals highlighted that communication effectiveness will likely increase if children have the literacy skills to access text-to-speech systems, as it is much quicker than searching for symbols. Tönsing et al. (2019) also advocated for increased literacy programmes for children with expressive communication difficulties. As speech synthesis systems, such as text-to-speech, offer children the ability to communicate independently, we are one step closer to countering the stigma and assumptions about a child with expressive communication difficulties' perceived level of intelligence (Ripat et al., 2019).

Limitations & Future Directions

Due to the small sample size and the methodology selected, the findings cannot be generalised to all AAC stakeholders in South Africa. As the syntax and morphology of African languages differs greatly from English, the WER results for isiXhosa should be interpreted cautiously. As an additional limitation, there are limited AAC devices/apps which are currently equipped to load synthetic African voices. Future research should consider gathering stakeholders' views about the selection of language-specific core vocabulary for AAC systems and researchers need to develop core vocabulary for each language. This core vocabulary could be paired with the synthetic voices, to form a functioning speech-generating system. It would also be important to gather perspectives from a larger group of children with expressive communication difficulties surrounding the applicability of high-tech AAC use.

Conclusions & Implications

Child AAC users are unable to explore the world of options that high-tech AAC offers them without the support of their communication partners. This study presented the perspectives of AAC stakeholders who reside in South Africa, a low- and middle-income country. Stakeholders from low- and middle-income countries are interested in the development of synthetic voices in their home languages, and our research highlights that children would prefer to incorporate these voices on their high-tech devices, and adults would prefer them for their children, learners, and/or clients' devices, rather than relying on British or US English voices. With a speech-generating device incorporating a child's home language, children with expressive communication difficulties would finally be able to actively participate in class and do so with a voice that matches their age, gender, social and linguistic background. Providing a variety of synthetic voice options and offering novel voices to children in a manner that respects and aligns with their linguistic and cultural backgrounds is crucial. Therefore, the development of synthetic speech in under-resourced languages has the potential to support marginalized AAC communities all over the world.

Supplementary Table 1

The Children's Responses, using a Pictographic 3-Point Scale, to Questions about the Quality, Acceptability, and Utility of the Synthetic Speech

		YES 	NOT SURE 	NO 
1	Do you like the child voice?	English participant Afrikaans participant isiXhosa participant		
2	Do you think the voice sounds like you?	English participant		Afrikaans participant isiXhosa participant
3	Does the language sound right?	English participant Afrikaans participant isiXhosa participant		
4	Do you think you would use this voice to help you talk at school?	English participant Afrikaans participant isiXhosa participant		
5	Do you think you would use this voice to help you talk at home?	English participant Afrikaans participant isiXhosa participant		
6	Will people be able to understand you	English participant Afrikaans participant isiXhosa participant		

	if you use this voice?		
7	Between the child and adult voice, would you prefer to use the adult voice?	isiXhosa participant	English participant Afrikaans participant
8	Between the child and adult voice, would you prefer to use the child voice?	English participant Afrikaans participant	isiXhosa participant
9	Do you want to tell me something else about the voice?		English participant Afrikaans participant isiXhosa participant

4. Discussion

4.1. General Discussion

This PhD study progressed through three distinct phases: first, reviewing speech synthesis technology worldwide, then exploring stakeholder perspectives about the barriers and facilitators to implementing AAC in LMICs, and finally, leading to the creation and evaluation of synthetic child speech in three under resourced South African languages. The evidence gathered from Phase 1a's scoping review highlighted the significant advancements achieved in speech synthesis technology in recent years. This progress has facilitated the development of synthetic speech with a remarkable level of naturalness, closely approaching the authentic pronunciation and intonation found in genuine speech (Terblanche et al., 2021; Wang et al., 2020; Wester et al., 2015). For example, it was found that a neural-network-based text-to-speech system, Tacotron 2, has gained popularity for its perceptual naturalness (Wang et al., 2017) and is therefore frequently chosen to generate synthetic speech. The widespread availability of open-source tools, such as Tacotron 2, has further normalised access to these technologies. Despite this, the scoping review highlighted a persistent challenge, especially pronounced when considering the communication needs of young children: the development of child speech synthesis.

The difficulty in collecting child speech data lies in its often inconsistent and imperfect recordings, and young children's speech presents with unique complexities, including acoustic variability, disfluencies, and articulatory errors (Govender et al., 2015). Given the challenges in collecting child speech data, the scoping review found that to obtain sufficient training data for synthetic child speech, researchers should consider leveraging both adult and child speech data (Begnum et al., 2012; Hagen et al., 2009; Karhila et al., 2012; Qian et al., 2016). Findings indicated that increasing the amount of training or adaptation data is associated with improved intelligibility, even when the data is imperfect (Govender et al., 2015; Hasija et al., 2021; Kumar & Surendra, 2011; Shivakumar & Georgiou, 2020). Moreover, several other considerations came to the forefront when contemplating the development of child speech synthesis. These include using average-voice models, the refinement of adaptation techniques for individual speaker targets, and the implementation of effective transform functions to ensure natural and intelligible synthesised speech. In summary, this scoping review provided insights into the challenges and considerations in developing child speech synthesis, emphasising the importance of language diversity, selecting appropriate synthesis systems and careful data collection.

Phase 1b aimed to understand the implementation of AAC systems within a more localised context, focusing specifically on the AAC landscape in LMICs, particularly South Africa. The study explored stakeholders' perspectives on both low- and high-tech AAC systems for children with CCN. Previous research highlights the crucial role communication partners play in effective AAC implementation (Jette et al., 2017; Kent-Walsh et al., 2015; Moorcroft et al., 2019b). This role is even more significant

in under-resourced settings, where challenges are compounded by socio-economic, cultural, and systemic factors. A major challenge identified in Phase 1b is the issue of reduced support and training for AAC implementation. While both high-income countries and LMICs face support issues (Baxter et al., 2012), LMICs often experience these challenges more severely due to a shortage of personnel and resources (van Niekerk et al., 2019), potentially making AAC service delivery unsustainable in these contexts. This finding is consistent with research from other developing countries, such as Kenya (Gona et al., 2014). Similarly, educating caregivers about AAC is a challenge in both contexts (Moorcroft et al., 2019a), but in Phase 1b, it was found that reduced caregiver training and education is often exacerbated in LMICs by factors such as caregivers' lack of transport, limited time off work and limited access to technology and/or basic resources. Additionally, while high staff turnover due to burnout is a global issue, Phase 1b found it to be prominent in LMICs due to high caseloads and a lack of professional support, as supported by research in Botswana and Egypt (Mukhopadhyay & Nwaogu, 2009; Wormnaes & Malek, 2004). In addition, language and code-switching issues further complicate AAC provision. While high-income countries also face a scarcity of AAC resources in local languages, LMICs experience even greater challenges due to higher linguistic diversity and fewer relevant resources (Amery et al., 2022; Baxter et al., 2012; Tönsing et al., 2018). The lack of locally relevant synthetic voices in LMICs is more pronounced due to limited technological development in speech synthesis for local languages (Tönsing et al., 2019).

Phase 1b found that certain challenges are unique to LMICs. Despite progressive disability policies in South Africa, accessibility to high-tech AAC devices remains limited (van Niekerk et al., 2019). Phase 1b highlighted the high risk of crime associated with expensive devices and cultural beliefs that stigmatise children with special needs, reducing device usability outside the school environment and hindering consistent communication support. Negative cultural perceptions of disabilities, which impact the acceptance of AAC by family members, were also observed in Kenya, another LMIC (Gona et al., 2014). Device affordability is another significant challenge in LMICs, where the high cost of AAC devices and software is prohibitive for many families (van Niekerk et al., 2019; Visagie et al., 2020). By identifying these practical issues faced by professionals and caregivers, Phase 1b shed light on the complex landscape of AAC implementation in a child's primary settings, including home and school. The proposed strategies from Phase 1b, such as strategic partnerships and evidence-based core-language AAC systems, highlight the need for solutions tailored to the realities of LMICs. The aim of Phase 1b was to gather insights into the requirements and difficulties of AAC implementation to ensure that future developments, as explored in Phase 2, are both practical and aligned with the unique contexts of people living in LMICs.

In Phase 2, the research shifted focus to the feasibility of creating synthetic child speech in under-resourced languages, even in scenarios with limited training data. Limited child speech data (less than

2 hours) were utilised with adult speech corpora (approximately 30 hours) to pre-train a Tacotron 2 model. Addressing a key proposal from the scoping review in Phase 1a, which advocated for the personalisation of synthetic voices to align with the unique vocal identity of children, and considering the constraints identified by stakeholders in Phase 1b, such as limited resources, Phase 2 successfully demonstrated that open-source speech synthesis technologies could be used to develop personalised synthetic child voices in different languages. The personalisation of synthetic child voices is recommended to enhance the efficacy and acceptance of AAC technologies (Mills et al., 2014; Ripat et al., 2019). By tailoring synthetic voices to reflect a child's vocal identity, including their language, sex and age, interactions with children using AAC devices become more natural and intuitive (Mills et al., 2014). This personalisation not only improves the voice quality but also makes AAC systems more accessible and responsive to individual needs (Amery et al., 2022; Jreige et al., 2009).

The method used in Phase 2 proves effective in creating child voices that authentically meet the communication needs of children with CCN in under-resourced languages. In Phase 2, 124 South African participants who spoke the language/s rated the naturalness of the synthetic voices using a 5-point Likert Scale. All the voices were rated as more natural than not. However, it was found that the synthetic child voices were rated less natural compared to adult voices, consistent with previous research showing that adult speech generally achieves higher naturalness ratings due to its greater fluency and fewer inherent distortions (Govender & de Wet, 2016; Jain et al., 2022). Additionally, the implementation of warm start procedures using adult speech data was shown to enhance the quality of the synthesised child voices. This approach aligns with Phuong et al. (2021), who demonstrated that a warm start can significantly improve synthesis quality and reduce training time. Phase 2 found no significant performance difference between using warm start procedures once or twice, suggesting that the quality of adaptation data is more crucial than the number of warm starts. This result highlights the value of using high-quality, age-appropriate models to improve child speech synthesis. Phase 2 confirmed that even minimal amounts of child speech data (as little as 5 minutes) could produce usable synthetic speech, with quality improvements observed with more data. This finding supports Shivakumar and Georgiou's (2020) conclusion that additional training data enhances synthesis quality. Overall, Phase 2 demonstrates the potential of modern open-source speech synthesis technologies, like Tacotron 2, in creating naturalistic child voices, even with limited data. This method opens new avenues for targeted technological interventions that cater specifically to the linguistic diversity of children in LMICs.

Phase 3 of the study aimed to evaluate and document the quality, acceptability, and utility of the synthetic voices developed in Phase 2, determining their viability as an addition to AAC in South Africa. It gathered perspectives from South African stakeholders on synthetic child speech, a topic not extensively covered in recent literature, particularly with input from diverse stakeholders from a LMIC

(Mukhopadhyay & Nwaogu, 2009; Tönsing et al., 2019; Tönsing & Dada, 2016). Stakeholders were generally accepting of the synthetic voices in different languages. The findings in Phase 3 showcase the preference amongst children and adults for incorporating these voices within high-tech devices, emphasising a departure from British or US English synthetic voices. Despite adult voices being rated higher than children's voices, similarly to the MOS ratings in Phase 2, none of the synthetic child voices received negative ratings on the Likert scale in terms of overall impression, pleasantness, naturalness, understandability, and similarity to real speakers. One reason for the lower perceived quality of synthetic children's voices was the unusual prosody caused by the slower speech rate and expressivity of the donor children. Participants may have attributed typical child speech characteristics like mid-sentence breaks, minor mispronunciations, and hesitations to the synthesis system rather than recognising them as natural elements of child speech (Jain et al., 2022). Nonetheless, the relatability and alignment with a child's vocal identity are crucial, and therefore synthesised child speech will frequently contain some child-like articulatory patterns (Begnum et al., 2012; Jain et al., 2022).

The English child voice, with its strong dialectal influence, was perceived as the poorest quality and least intelligible, unlike the Afrikaans and isiXhosa voices with less prominent dialects. This preference for standard dialects from the adult participants might be due to greater exposure, preconceived notions about synthetic voices, or the perceived prestige of standard dialects (Gates & Ilbury, 2019; Perry et al., 2018). Despite the preference for standard dialects, synthetic speech with different dialects empowers children with CCN as it offers them a choice. This aspect of choice is crucial as it reflects the broader significance of one's voice in fostering identity and belonging. Children may feel a stronger sense of belonging when they can select voices that reflect their own dialect and language. This highlights the need for AAC resources in children's home languages to support academic engagement and cultural preservation, which is especially important in a multilingual country like South Africa (Ndlangamandla, 2010; Tönsing et al., 2019). With that in mind, the children with CCN gave their opinions about their synthetic voices. The children were generally positive about the voices, but only the English child felt the synthetic voice matched their own. The isiXhosa child preferred the adult voice, possibly due to their age, while the younger Afrikaans and English children preferred the child voices. All children agreed that the language sounded appropriate and would be understood at school and home.

Adult participants in Phase 3 noted that increased AAC training in the classroom could help children thrive with these devices, while inconsistent use of AAC and poor transitioning to AAC systems with larger vocabularies remain barriers to use. Participants also suggested that bullying and teasing can lead to AAC abandonment, but peer exposure to AAC may reduce this. Peers can model AAC use, assist with troubleshooting, and improve their own language and digital literacy skills (Finke et al., 2009). Professionals in Phase 3 also highlighted that low-tech AAC systems in classrooms encourage early use

and ease transitioning to high-tech systems. By presenting the perspectives of AAC stakeholders in South Africa, Phase 3 highlights the importance of developing and evaluating synthetic voices in under-resourced home languages. In summary, this PhD research project not only advances the theoretical understanding of synthetic child speech development but also attempts to understand the unique challenges faced by stakeholders in LMICs, providing meaningful and contextually relevant benefits for AAC users and their communication partners.

4.2. Implications of the Findings

This PhD research project has both theoretical and applied implications. Considering the theoretical implications, this project focuses on using pre-trained adult models to create personalised synthetic child voices for children with CCN, in three South African languages, thereby bridging the fields of computational linguistics and AAC. By creating a connection between these two fields, the research expands and generates new knowledge, showcasing the feasibility of generating synthetic child speech in under-resourced languages, and demonstrating the potential of using open-source technological advancements to address communication needs for children with CCN across diverse linguistic and socio-economic contexts. Additionally, findings from this study play a pivotal role in documenting the process to create synthetic child speech in under-resourced languages, creating a viable method for researchers to replicate the process and develop synthetic child speech in other languages, using primarily open source tools, which are readily available and low-cost. It also brings attention to critical issues in resource-limited settings and underscores the importance of cultural relevance in enhancing communication experiences, specifically those of children from marginalised groups, such as those with communication disabilities.

Looking at the applied implications, gathering perspectives from stakeholders provides valuable insights into the acceptability of synthetic speech, crucial for guiding practical applications. The findings highlight stakeholder acceptance of synthetic voices in SAE, Afrikaans, and isiXhosa, affirming the practical viability of implementing synthetic voices in LMICs. In addition, this research advocates for inclusive AAC solutions, emphasising the significance of locally-sounding synthetic voices, promoting accessibility. Finally, in a tangible application, the research provided iPads with synthetic voices to three children with CCN, enabling them to communicate independently for the first time, and in their home languages. This not only benefits the children but also enhances the effectiveness of communication for their communication partners at home and at school. The demonstrated method for generating natural-sounding, personalised synthetic child voices has the potential to empower more children with CCN in other under-resourced settings, allowing active participation in class and providing voices that match their individual identities and cultures. The dual implications – theoretical and applied – significantly contribute to the ongoing dialogue on tailoring

communication technologies to specific cultural and linguistic contexts while actively addressing the needs of individuals with CCN, and supporting marginalised AAC communities globally.

4.3. Strengths, Limitations of the study and Future Research Directions

This PhD study explores a novel area by applying modern open-source speech synthesis technologies to synthetic child voices in under-resourced languages, addressing a significant gap in the literature. In contrast to the common underrepresentation and lack of multi-perspective input from stakeholders in LMICs (Baxter et al., 2012), this PhD research project actively includes the diverse perspectives of stakeholders, including teachers, SLTs, caregivers, and children with CCN in South Africa, an LMIC. While caregivers and SLTs play a crucial role in decision-making for a child, this study also recognises the importance of including children's opinions when selecting synthetic voices. Furthermore, the project's use of open-source speech synthesis software ensures feasibility and allows other researchers to replicate the process for additional under-resourced languages. This research highlights the significance of offering diverse synthetic voice options for children and showcases the importance of introducing novel voices for high-tech AAC in a way that respects and aligns with their linguistic and cultural backgrounds.

Despite the strengths of the study, several limitations must be acknowledged. During Phase 1a, relevant articles might have been excluded, as the selected databases may not have included all pertinent studies. The utilisation of alternative databases or varied search terminology could have unveiled additional relevant literature. The scoping review also refrained from critically appraising individual sources, as it focused on identifying available evidence rather than evaluating its quality. In Phase 1b and Phase 3, the focus group studies encountered limitations due to a small sample size and a chosen methodology, namely focusing on stakeholders who have frequent interactions with the target children, which may have restricted generalisability to all AAC stakeholders in LMICs. Recruiting families from low socio-economic backgrounds proved challenging, with financial constraints and childcare issues hindering participation. While all participants agreed to conduct focus group discussions in English without a translator, language limitations might have restricted in-depth conversations for some. In Phase 2, the use of the established Tacotron 2 architecture for synthetic speech creation was not necessarily a novel approach. However, employing modern text-to-speech technologies in under-resourced languages for children with CCN in LMICs is a novel focus that lacks adequate attention in the existing literature. Moreover, as the subjective quality assessments used in Phase 2 were conducted online, participants may have potentially experienced some technical issues, and the limited number of participants, especially in the isiXhosa category, implies caution in generalising the results. Another significant limitation is the current lack of AAC devices/apps compatible with the African synthetic voices developed during this phase of the research. The study was mindful of ensuring that voices were

culturally relevant to the children with CCN, acknowledging that language and accent are central to identity. However, the researcher also recognises that cultural variations within the language groups, such as accent or speech patterns specific to certain regions, communities, or socioeconomic groups, could offer a more nuanced and representative voice in future work. Additionally, it is important to acknowledge that the perceptions of peers were not gathered in Phase 4, which would have been an interesting discussion point. Moreover, the study made use of decontextualised synthetic speech, and perceptions were never gathered by using the AAC in "real time conversations", which would be an important next step. Lastly, since this study was largely exploratory, the final limitation is that the synthetic child voices were only made for a small number of children.

Exploring future research avenues, especially in the field of AAC, and particularly speech synthesis technology, holds significant promise for advancing our understanding and enhancing AAC communication solutions for diverse populations. Firstly, future research should consider gathering the views of a broader spectrum of AAC stakeholders at other ecological levels and mapping the process of change through theory of change workshops. Secondly, while this PhD research demonstrated the viability of generating synthetic child voices in three under-resourced languages, further investigations are warranted to extend this capability to other under-resourced languages, and efforts should be directed towards the development of accompanying intervention guides. One would need to incorporate graphic symbols with cross-cultural readability into a system to support children who are illiterate, or one could simply utilize a simple text-to-speech system in each language. However, using a system without symbols may limit its efficacy for many of the South African children with special needs, due to the low literacy levels amongst this population. Thirdly, broader perspectives from a diverse range of AAC end-users, including children with CCN, are pivotal in evaluating the suitability of high-tech AAC applications. Gathering stakeholders' insights on language-specific core vocabulary selection for AAC systems is also crucial, especially for under-resourced languages. Since there are already established speech-generating systems, one could either incorporate this language-specific core vocabulary and corresponding synthetic voice into compatible devices or create a new app to facilitate this integration. It may also be worthwhile for researchers to explore the possibility of code-switching in speech synthesis, and to investigate the capability for synthetic voices to evolve and age in tandem with the child's natural aging process. It also may be interesting to pre-train a model with child data rather than adult data. This would necessitate the creation of substantial corpora of child speech data in diverse languages, which also stands as a potential avenue for future research. Lastly, language and culture are deeply interconnected, and voice characteristics can vary not only based on linguistic elements but also on cultural nuances, including regional accents, speech patterns, and social identity. Future studies could benefit from developing synthetic voices that account for a broader range of cultural variations within each language, ensuring that voices are more representative of the diverse cultural identities within the respective language communities.

4.4. Conclusion

The significance of access to speech synthesis on AAC devices for children, particularly those with CCN, is profound. This PhD research project, has made a meaningful contribution to the field of child speech synthesis, and advanced our understanding of contextually relevant AAC solutions in LMICs. The unique differences in child speech, such as acoustic variability and articulatory errors, make the development of synthetic child voices a challenging yet interesting area of research. The potential implications of quality child speech synthesis extend beyond technological advancements in the field of speech synthesis, and holds real value as a tool to improve access and participation of children with CCN, particularly in educational contexts. In the LSEN classroom, where multiple children may benefit from synthetic voices, individualised vocal identities could facilitate speaker differentiation, potentially enhancing technology adoption rates, increasing social interaction among children and fostering positive academic trajectories. Synthetic child speech allows children with CCN to participate independently, increasing opportunities for educational, social, and emotional development. This PhD project has significantly enhanced communication and participation opportunities for three children who were previously unable to communicate using their natural voice. By addressing the unique needs of children in under-resourced languages, and incorporating stakeholder perspectives, this PhD research project has advocated for social justice and equity, ensuring that all children, regardless of the language they speak, have access to high-quality speech synthesis. This research reinforces the transformative potential of speech synthesis in fostering communication equality and ensuring that every child, regardless of linguistic or socio-economic background, has a voice that resonates with their unique identity and culture.

5. References

References marked with an asterisk indicate studies identified as a result of the scoping review search in Phase 1.

American Speech-Language-Hearing Association. (2022). *Augmentative and alternative communication (AAC)*. Augmentative and Alternative Communication. https://www.asha.org/NJC/AAC/

Amery, R., Thirumanickam, A., Barker, R., Lowell, A., Theodoros, D., & Raghavendra, P. (2022). Developing Augmentative and Alternative Communication Systems in Languages Other Than English: A Scoping Review. American Journal of Speech-Language Pathology, 31(6), 2900–2919. https://doi.org/10.1044/2022_AJSLP-21-00396

Anastasiou, D., & Kauffman, J. M. (2013). The Social model of disability: Dichotomy between impairment and disability. *Journal of Medicine and Philosophy*, *38*, 441–459. https://doi.org/10.1093/jmp/jht026

Anumanchipalli, G. K., & Black, A. B. (2010). Adaptation techniques for speech synthesis in under-resourced languages. *Spoken Language Technologies for Under-Resourced Languages*.

Attainment Company (2011) GoTalks. Available at: https://www.attainmentcompany.com/technology/gotalks.

Baxter, S., Enderby, P., Evans, P., & Judge, S. (2012). Barriers and facilitators to the use of high-technology augmentative and alternative communication devices: A systematic review and qualitative synthesis. *International Journal of Language & Communication Disorders*, *47*(2), 115–129. https://doi.org/10.1111/j.1460-6984.2011.00090.x

*Begnum, M., Flatebø Hoelseth, S., Johnsen, B., & Hansen, F. (2012). A Child's Voice. *Norsk Informatikkonferanse (NIK) Conference (9-12 November)*, 165–176.

Beukelman, D. R., & Light, J. C. (2020). *Augmentative & Alternative Communication: Supporting Children and Adults with Complex Communication Needs* (5th ed.). Brookes Publishing.

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, *56*, 85–100. https://doi.org/10.1016/j.specom.2013.07.008

Boesch, M., & Da Fonte, M. (2019). *Effective Augmentative and Alternative Communication Practices: A Handbook for School-Based Practitioners*. Routledge.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Braun, V., & Clarke, V. (2021). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, *18*(3), 328–352. https://doi.org/10.1080/14780887.2020.1769238

Bruno, J. (2010). *Test of Aided Symbol Performance*. DynaVox Mayer-Johnson.

Christensen, R. H. B. (2022). *Ordinal—Regression models for ordinal data* [R package version 2022.11-16]. https://CRAN.R-project.org/package=ordinal.

Clark, R. A. J., Podsiadło, M., Fraser, M., Mayo, C., & King, S. (2007). Statistical analysis of the Blizzard Challenge 2007 listening test results. *The Blizzard Challenge 2007*.

Colquhoun, H. L., Levac, D., O'Brien, K. K., Straus, S., Tricco, A. C., Perrier, L., Kastner, M., & Mohere, D. (2014). Scoping reviews: Time for clarity in definition, methods, and reporting. *Journal of Clinical Epidemiology*, *64*, 1291–1294. http://dx.doi.org/10.1016/j.jclinepi.2014.03.013

*Cosi, P. (2009). On the development of matched and mismatched Italian children's speech recognition systems. *Interspeech*, 540–543.

*Cosi, P. (2015). *A KALDI-DNN-based ASR system for Italian: Experiments on Children Speech*. 1–5.

*Cosi, P., Nicolao, M., Paci, G., Sommavilla, G., & Tesser, F. (2014). Comparing open source ASR toolkits on Italian children speech. *Fourth Workshop on Child Computer Interaction (WOCCI)*.

Creer, S. (2009). *Personalising synthetic voices for individuals with severe speech impairment* [PhD thesis]. University of Sheffield.

Creer, S., Cunningham, S., Green, P., & Yamagishi, Y. (2013). Building personalised synthetic voices for individuals with severe speech impairment. *Computer Speech and Language*, *27*(6), 1178–1193. https://doi.org/10.1016/j.csl.2012.10.001

Creswell, J., W., & Clark, V., L. P. (2007). Choosing a mixed methods design. In *Designing and conducting mixed methods research* (pp. 58–88). SAGE Publications.

*Cui, X., & Alwan, A. (2006). Adaptation of children's speech with limited data based on formant-like peak alignment. *Computer Speech and Language*, *20*, 400–419.

Da Fonte, M., & Boesch, M. C. (2019). *Effective Augmentative and Alternative Communication Practices: A Handbook for School-Based Practitioners*. Routledge.

Dada, S., Flores, C., Bastable, K., & Schlosser, R. W. (2021). The effects of augmentative and alternative communication interventions on the receptive language skills of children with developmental disabilities: A scoping review. *International Journal of Speech-Language Pathology*, *23*(3), 247–257. https://doi.org/10.1080/17549507.2020.1797165

Dada, S., Horn, T., Samuels, A., & Schlosser, R. W. (2016). Children's attitudes toward interaction with an unfamiliar peer with complex communication needs: Comparing high- and low-technology devices. *Augmentative and Alternative Communication*, *32*(4), 305–311.

Dada, S., Murray, J., & Smith, M. (2022). Augmentative and Alternative Communication in Underserved or Unserved Populations. In S. Levey & P. Enderby (Eds.), *The Unserved: Addressing the needs of those with communication disorders* (pp. 109–118). J & R Press Ltd.

de Wet, F., Van der Walt, W., Dlamini, N., & Govender, A. (2017). Building synthetic voices for under-resourced languages: The feasibility of using audiobook data. *Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*.

Donohue D and Bornman J (2014) The challenges of realising inclusive education in South Africa. *South African journal of education* 34(2).

*Drager, K., & Finke, E. (2012). Intelligibility of Children's Speech in Digitized Speech. *AAC: Augmentative & Alternative Communication*, *28*(3), 181–189.

*Drager, K., Clark-Serpentine, E., Johnson, K., & Roeser, J. (2006). Accuracy of Repetition of Digitized and Synthesized Speech for Young Children in Background Noise. *American Journal of Speech-Language Pathology*, *15*(2), 155–164.

Drager, K., Justad, K., & Gable, K. (2004). Telephone communication: Synthetic and dysarthric speech intelligibility and listener preferences. *Augmentative and Alternative Communication*, *20*(2), 103–112.

Drager, K., Light, J., & McNaughton, D. (2010). Effects of AAC interventions on communication and language for young children with complex communication needs. *Journal of Pediatric Rehabilitation Medicine: An Interdisciplinary Approach*, *3*, 303–310. https://doi.org/10.3233/PRM-2010-0141

*Drager, K., Reichle, J., & Pinkoski, C. (2010). Synthesized Speech Output and Children: A Scoping Review. *American Journal of Speech-Language Pathology*, *19*, 259–273.

*Fainberg, J., Bell, P., Lincoln, M., & Renals, S. (2016). *Improving Children's Speech Recognition through Out-of-Domain Data Augmentation*. Interspeech, San Francisco, USA (September 8–12).

Finke, E. H., McNaughton, D. B., & Drager, K. D. R. (2009). "All Children Can and Should Have the Opportunity to Learn": General Education Teachers' Perspectives on Including Children with Autism Spectrum Disorder who Require AAC. *Augmentative and Alternative Communication*, *25*(2), 110–122. https://doi.org/10.1080/07434610902886206

*Fringi, E., Lehman, J., & Russel, M. (2015). Evidence of phonological processes in automatic recognition of children's speech. *Sixteenth Annual Conference of the International Speech Communication Association*.

Ganya, W., Kling, S., & Moodley, K. (2016). Autonomy of the child in the South African context: Is a 12 year old of sufficient maturity to consent to medical treatment? *BMC Medical Ethics*, *17*(1), 66–66. https://doi.org/10.1186/s12910-016-0150-0

Ganz, J., Hong, E., & Goodwyn, F. (2012). Effectiveness of the PECS Phase III app and choice between the app and traditional PECS among preschoolers with ASD. *Research in Autism Spectrum Disorders*, *7*(8), 973–983.

Gates, S. M., & Ilbury, C. (2019). Standard language ideology and the non-standard adolescent speaker. In C. Wright, L. Harvey, & J. Simpson (Eds.), *Voices and Practices in Applied Linguistics* (pp. 109–126). White Rose University Press; JSTOR.

*Gerosa, M., Giuliani, D., & Brugnara, F. (2007). Acoustic variability and automatic recognition of children's speech. *Speech Communication*, *49*(10–11), 847–860.

*Gerosa, M., Giuliani, D., & Brugnara, F. (2009). Towards age-independent acoustic modeling. *Speech Communication*, *51*(6). https://doi.org/10.1016/j.specom.2009.01.006

*Gerosa, M., Giuliani, D., Narayanan, S., & Potamianos, A. (2009). A review of ASR technologies for children's speech. *2nd Workshop on Child, Computer and Interaction (Interspeech ICSLP)*, 1–8.

*Ghai, S., & Sinha, R. (2009). Exploring the role of spectral smoothing in context of children's speech recognition. *Interspeech (6-10 September)*, 1607–1610.

*Ghai, S., & Sinha, R. (2010a). *Analyzing pitch robustness of PMVDR and MFCC features for children's speech recognition*. 1–5.

*Ghai, S., & Sinha, R. (2010b). Enhancing children's speech recognition under mismatched condition by explicit acoustic normalization. *Interspeech*.

*Ghai, S., & Sinha, R. (2010c). Exploring the effect of differences in the acoustic correlates of adults' and children's speech in the context of automatic speech recognition. *EURASIP Journal on Audio Speech and Music Processing*, 1–15. https://doi.org/10.1155/2010/318785

*Ghai, S., & Sinha, R. (2015). Pitch adaptive MFCC features for improving children's mismatched ASR. *International Journal of Speech Technology*, *18*, 489–503. https://doi.org/10.1007/s10772-015-9291-7

*Giuliani, D., & BabaAli, B. (2015). *Large Vocabulary Children's Speech Recognition with DNN-HMM and SGMM Acoustic Modeling*. 1635–1639.

*Giuliani, D., Gerosa, M., & Brugnara, F. (2006). Improved automatic speech recognition through speaker normalization. *Computer Speech and Language*, *20*, 107–123.

Global Research on Developmental Disabilities Collaborators (2018). Developmental disabilities among children younger than 5 years in 195 countries and territories, 1990-2016: a systematic

analysis for the Global Burden of Disease Study 2016. *The Lancet. Global health*, *6*(10), e1100–e1121. https://doi.org/10.1016/S2214-109X(18)30309-7

Gona, J. K., Newton, C. R., Hartley, S., & Bunning, K. (2014). A home-based intervention using augmentative and alternative communication (AAC) techniques in rural Kenya: What are the caregivers' experiences? *Child: Care, Health and Development*, *40*(1).

Gorenflo, C. W., Gorenflo, D. W., & Santer, S. A. (1994). Effects of synthetic voice output on attitudes toward the augmented communicator. *Journal of Speech and Hearing Research*, *37*, 64–68.

*Govender, A., & de Wet, F. (2016). Objective measures to improve the selection of training speakers in HMM-based child speech synthesis. *Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*.

*Govender, A., Nouhou, B., & De Wet, F. (2015). *HMM Adaptation for child speech synthesis using ASR data*. 2015 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), Port Elizabeth, South Africa.

*Gray, S., Willett, D., Lu, J., Pinto, J., Maergner, P., & Bodenstab, N. (2014). Child automatic speech recognition for US English: Child interaction with living-room-electronic-devices. *Fourth Workshop on Child Computer Interaction (WOCCI)*, 21–26.

Gutkin, A., Ha, L., Jansche, M., Pipatsrisawat, K., & Sproat, R. (2016). TTS for Low Resource Languages: A Bangla Synthesizer. *LREC*, 2005–2010.

*Hagen, A., Pellom, B., & Hacioglu, K. (2009). *Generating synthetic children's acoustic models from adult models*. 77–80.

Hamadeh, N., Van Rompaey, C., & Metreau, E. (2023). World Bank Group country classifications by income level for FY24. *Data Blog*. https://blogs.worldbank.org/en/opendata/new-world-bank-group-country-classifications-income-level-fy24

*Hasija, T., Kadyan, V., & Guleria, K. (2021). *Out Domain Data Augmentation on Punjabi Children Speech Recognition using Tacotron*. International Conference on Mathematics and Artificial Intelligence (ICMAI 2021) (March 19 - 21, 2021), Chengdu, China. https://doi.org/10.1088/1742-6596/1950/1/012044

Hoover, J., Reichle, J., Van Tasell, D., & Cole, D. (1987). The intelligibility of synthesized speech: Echo II versus Votrax. *Journal of Speech and Hearing Research*, *30*(3), 425–431. https://doi.org/10.1044/jshr.3003.425

Ito, K., & Johnson, L. (2017). *The LJ speech dataset*. https://keithito.com/LJ-Speech-Dataset/

*Jacob, A., & Mythili, P. (2008). Developing a Child Friendly Text-to-Speech System. *Advances in Human-Computer Interaction*. https://doi.org/10.1155/2008/597971

Jain, R., Yiwere, M. Y., Bigioi, D., Corcoran, P., & Cucu, H. (2022). A text-to-speech pipeline, evaluation methodology, and initial fine-tuning results for child speech synthesis. *IEEE Access*, *10*, 47628-47642. https://doi.org/10.1109/ACCESS.2022.3170836

Jette, A. M., Spicer, C. M., & Flaubert, J. L. (Eds.). (2017). Augmentative and Alternative Communication and voice products and technologies. In *The promise of assistive technology to enhance activity and work participation* (pp. 209–310). The National Academies Press. https://doi.org/10.17226/24740

*Jia, N., Zheng, C., & Sun, W. (2020). *Speech synthesis of children's reading based on cycleGAN model*. International Symposium on Electronic Information Technology and Communication Engineering (ISEITCE) (June 19-21, 2020), Jinan, China.

*Jreige, C., Patel, R., & Bunnell, H. T. (2009). VocaliD: Personalizing text-to-speech synthesis for individuals with severe speech impairment. *Assets '09*, 259–260.

*Karhila, R., Sanand, D. R., Kurimo, M., & Smit, P. (2012). Creating synthetic voices for children by adapting adult average voice using stacked transformations and VTLN. *ICASSP 2012*. 37th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Kyoto, Japan.

Kathard H, Ramma L, Pascoe M, et al. (2011) How can speech-language therapists and audiologists enhance language and literacy outcomes in South Africa? (And why we urgently need to. *South African Journal of Communication Disorders* 58(2).

Kent-Walsh, J., Murza, K., Malani, M. D., & Binger, C. (2015). Effects of communication partner instruction on the communication of individuals using AAC: A meta-analysis. *Augmentative and Alternative Communication*, 1–14. https://doi.org/10.3109/07434618.2015.1052153

Kirkhart, K., & Hopson. (2010). *"Strengthening Evaluation Through Cultural Relevance and Cultural Competence." Invited workshop.* American Evaluation Association/Centers for Disease Control Summer Institute, June 13-16, 2010, Atlanta.

*Koul, R., & Clapsaddle, K. (2006). Effects of repeated listening experiences on the perception of synthetic speech by individuals with mild-to-moderate intellectual disabilities. *AAC: Augmentative and Alternative Communication*, *22*(2), 112–122.

*Koul, R., & Hester, K. (2006). Effects of Repeated Listening Experiences on the Recognition of Synthetic Speech by Individuals With Severe Intellectual Disabilities. *Journal of Speech, Language, and Hearing Research*, *49*, 47–57.

*Kumar, J. V., & Surendra, A. K. (2011). Statistical Parametric Approach for Child Speech Synthesis using HMM-Based System. *International Journal of Computer Science & Technology*, *2*(1), 149–152.

Leedy, P., & Ormrod, J. (2013). *Practical research: Planning and design* (10th ed.). Pearson.

Leonet O, Orcasitas-Vicandi M, Langarika-Rocafort A, et al. (2022) A systematic review of augmentative and alternative communication interventions for children aged from 0 to 6 Years. *American Speech-Language hearing Association (ASHA)* 53: 894–920.

Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: Advancing the methodology. *Implementation Science*, *5*(69). http://www.implementationscience.com/content/5/1/69

Liamputtong P (2011) *Focus Group Methodology: Introduction and History*. SAGE Publications Ltd.

*Liao, H., Pundak, G., Siohan, O., Carroll, M., Coccaro, N., Jiang, Q., Sainath, T., Senior, A., Beaufays, F., & Bacchiani, M. (2015). Large vocabulary automatic speech recognition for children. *Interspeech (6-10 September)*, 1611–1615.

Light, J., McNaughton, D., & Caron, J. (2019). New and emerging AAC technology supports for children with complex communication needs and their communication partners: State of the science and future research directions. *Augmentative and Alternative Communication*, 1–16. https://doi.org/10.1080/07434618.2018.1557251

Louw, A., & Schlünz, G. (2016a). *Lwazi III Afrikaans TTS Corpus* (ISLRN 605-808-477-011-9; 1st ed.) [Dataset]. Meraka Institute, CSIR. https://hdl.handle.net/20.500.12185/266

Louw, A., & Schlünz, G. (2016b). *Lwazi III English TTS Corpus* (ISLRN 266-833-874-480-1; Vol. 1) [Dataset]. Meraka Institute, CSIR. https://hdl.handle.net/20.500.12185/267

Louw, A., & Schlünz, G. (2016c). *Lwazi III isiXhosa TTS Corpus* (ISLRN 038-391-782-117-6; Vol. 1) [Dataset]. Meraka Institute, CSIR. https://hdl.handle.net/20.500.12185/268

*Matassoni, M., Falavigna, D., & Giuliani, D. (2016). DNN adaptation for recognition of children speech through automatic utterance selection. *Spoken Language Technology Workship (IEEE)*, 644–651.

*Matassoni, M., Gretter, R., Falavigna, D., & Giuliani, D. (2018). Non-native children speech recognition through transfer learning. *Acoustics, Speech and Signal Processing (ICASSP)*, 6229–6233.

*Metallinou, A., & Cheng, J. (2014). Using deep neural networks to improve proficiency assessment for children english language learners. *Interspeech (14-18 September)*, 1468–1472.

*Mills, T., Bunnell, T., & Patel, R. (2014). Towards personalized speech synthesis for augmentative and alternative communication. *Augmentative and Alternative Communication*, *30*(3), 226–236. https://doi.org/10.3109/07434618.2014.924026

Moorcroft, A., Scarinci, N., & Meyer, C. (2019a). A systematic review of the barriers and facilitators to the provision and use of low-tech and unaided AAC systems for people with complex communication needs and their families. *Disability and Rehabilitation: Assistive Technology*, *14*(7), 710–731. https://doi.org/10.1080/17483107.2018.1499135

Moorcroft, A., Scarinci, N., & Meyer, C. (2019b). Speech pathologist perspectives on the acceptance versus rejection or abandonment of AAC systems for children with complex communication needs. *Augmentative and Alternative Communication*, *35*(3), 193–204. https://doi.org/10.1080/07434618.2019.1609577

*Mousa, A. (2011). Speech segmentation in synthesized speech morphing using pitch shifting. *International Arab Journal of Information Technology*, *8*(2), 221–226.

Mukhopadhyay, S., & Nwaogu, P. (2009). Barriers to teaching non-speaking learners with intellectual disabilities and their impact on the provision of augmentative and alternative communication. *International Journal of Disability, Development and Education*, *56*, 349–362. http://hdl.handle.net/10311/532

*Murphy, A., Yanushevskaya, I., Chasaide, A. N., & Gobl, C. (2020). *Testing the GlórCáil system in a speaker and affect voice transformation task*. 950–954.

Nathanson, E. (2017). Native voice, self-concept and the moral case for personalized voice technology. *Disability and Rehabilitation*, *39*(1), 73–81. https://doi.org/10.3109/09638288.2016.1139193

Ndlangamandla, S. C. (2010). Multilingualism in desegregated schools: Learners' use of and views towards African languages. *Southern African Linguistics and Applied Language Studies*, *28*(1), 61–73. https://doi.org/10.2989/16073614.2010.488444

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—A web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 210. https://doi.org/10.1186/s13643-016-0384-4

Pascoe, M., & Norman, V. (2011) Contextually relevant resources in speech-language therapy and audiology in South Africa – are there any?*. *South African Journal of Communication Disorders* 58(1): 2–5.

Perry, L. K., Mech, E. N., MacDonald, M. C., & Seidenberg, M. S. (2018). Influences of speech familiarity on immediate perception and final comprehension. *Psychonomic Bulletin & Review*, *25*(1), 431–439. https://doi.org/10.3758/s13423-017-1297-5

Phuong, P. N., Quang, C. T., Do, Q. T., & Luong, M. C. (2021). *A study on neural-network-based text-to-speech adaptation techniques for Vietnamese.* 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques.

*Přibilová, A., & Přibil, J. (2006). Non-linear frequency scale mapping for voice conversion in text-to-speech system with cepstral description. *Speech Communication*, *48*, 1691–1703.

Pucher, M., Schuchmann, G., & Fröhlich, P. (2009). Regionalized text-to-speech systems: persona design and application scenarios. In A. Esposito, A. Hussain, M. Marinaro, & R. Martone (Eds.), *Multimodal Signals: Cognitive and Algorithmic Issues* (pp. 216–222). Springer Berlin Heidelberg.

*Pucher, M., Toman, M., Schabus, D., Valentini-Botinhao, C., Yamagishi, J., Zillinger, B., & Schmid, E. (2015). *Influence of speaker familiarity on blind and visually impaired children's*

perception of synthetic voices in audio games. Interspeech (September 6-10, 2015), Dresden, Germany.

*Qian, M., McLoughlin, I., Guo, W., & Dai, L. (2016). *Mismatched Training Data Enhancement for Automatic Recognition of Children's Speech using DNN-HMM*. Chinese Spoken Language Processing (ISCSLP), 10th International Symposium.

*Qian, Y., Wang, X., Evanini, K., & Suendermann-Oeft, D. (2016). *Improving DNN-Based Automatic Recognition of Non-native Children's Speech with Adult Speech*. 40–44.

QSR International. (1998). *NVivo qualitative data analysis [Software]* (1.6.2) [Computer software]. https://qsrinternational.com/nvivo/nvivo-products/

R Core Team. (2019). *R: A language and environment for statistical computing. R Foundation for Statistical Computing* [Computer software]. https://www.R-project.org/

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

Ripat, J., Verdonck, M., Gacek, C., & McNicol, S. (2019). A qualitative metasynthesis of the meaning of speech-generating devices for people with complex communication needs. *Augmentative and Alternative Communication*, *35*(2), 69–79. https://doi.org/10.1080/07434618.2018.1513071

*Saheer, L., Yamagishi, J., Garner, P. N., & Dines, J. (2013). Combining vocal tract length normalization with hierarchical linear transformations. *IEEE Journal of Selected Topics in Signal Processing*, *8*(2), 262–272.

Schlosser, R., Sigafoos, J., Luiselli, J., Angermeier, K., Harasymowyz, U., Schooley, K., & Belfiore, P. (2007). Effects of synthetic speech output on requesting and natural speech production in children with autism: A preliminary study. *Research in Autism Spectrum Disorders*, *1*(2), 139–163.

Sefara, T. J., Mokgonyane, T. B., Manamela, M. J., & Modipa, T. I. (2019). HMM-based speech synthesis system incorporated with language identification for low-resourced languages. *International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*. https://doi.org/10.1109/ICABCD.2019.8851055

*Serizel, R., & Giuliani, D. (2014). *Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition*. 135–140. https://doi.org/10.1109/SLT.2014.7078563

*Serizel, R., & Giuliani, D. (2016). Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children. *Natural Language Engineering*, *23*(3), 325–350. https://doi.org/10.1017/S135132491600005X

*Shahnawazuddin, S., Dey, A., & Sinha, R. (2016). Pitch-adaptive front-end features for robust children's ASR. *Interspeech (8-12 September)*, 3459–3463.

*Shivakumar, P. G., & Georgiou, P. (2020). Transfer Learning from Adult to Children for Speech Recognition: Evaluation, Analysis and Recommendations. *Comput Speech Lang*, *63*(8), 1–46. https://doi.org/10.1016/j.csl.2020.101077

*Shivakumar, P. G., Potamianos, A., Lee, S., & Narayanan, S. (2014). Improving speech recognition for children using acoustic adaptation and pronunciation modeling. *Fourth Workshop on Child Computer Interaction (WOCCI)*, 15–19.

*Sinha, R., & Ghai, S. (2009). On the use of pitch normalization for improving children's speech recognition. *Interspeech*, 568–571.

South African Department of Education (2001) Education White Paper 6:Building an inclusive education and training system. Available at: https://www.education.gov.za/Resources/Legislation/WhitePapers.aspx.

South African Government. (1997). *Language in Education Policy Document (LiEP)*. Pretoria: Government Printers. https://www.education.gov.za/Portals/0/Documents/Policies/GET/LanguageEducationPolicy1997.pdf

Statistics SA. (2020). Child Poverty in South Africa: A Multiple Overlapping Deprivation Analysis. *Statistics SA*. www.statssa.gov.za

Statistics South Africa. (2022). *Census 2022*. Statistics South Africa. https://census.statssa.gov.za/assets/documents/2022/P03014_Census_2022_Statistical_Release.pdf

Sutton, S., Foulkes, P., Kirk, D., & Lawson, S. (2019). Voice as a design material: Sociophonetic inspired design strategies in Human-Computer Interaction. *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)*. https://doi.org/10.1145/3290605.3300833

Tashakkori, A., & Creswell, J. W. (2007). Editorial: The new era of mixed methods. *Journal of Mixed Methods Research*, *1*(1), 3–7. https://doi.org/10.1177/2345678906293

Tegler H, Pless M, Blom Johnson M, et al. (2019) Speech and language pathologists' perceptions and practises of communication partner training to support children's communication with high-tech speech generating devices. *Assistive Technology* 14(6): 581–589.

Terblanche, C., Harrison, P., & Gully, A. (2021). *Human spoofing detection performance on degraded speech*. 1738–1742. https://doi.org/10.21437/Interspeech.2021-1225

Terblanche, C., Harty, M., Pascoe, M., & Tucker, B. V. (2022). A situational analysis of current speech-synthesis systems for child voices: A scoping review of qualitative and quantitative Evidence. *Applied Sciences*, *12*(5623), 1–17. https://doi.org/10.3390/app12115623

Terblanche, C., Schnoor, T. T., Harty, M., & Tucker, B. V. (in press). The Development of Synthetic Child Speech in Three South African Languages. *Augmentative and Alternative Communication*. https://doi.org/10.1080/07434618.2024.2374312

Terre blanche M, Durrheim K and Painter D (eds) (2006) *Research in Practice: Applied Methods for the Social Sciences*. 2nd ed. University of Cape Town Press.

*Tong, R., Chen, N., & Ma, B. (2017). *Multi-Task Learning for Mispronunciation Detection on Singapore Children's Mandarin Speech*. 2193–2197.

*Tong, R., Wang, L., & Ma, B. (2017). *Transfer learning for children's speech recognition*. 2193–2197.

Tönsing, K., & Dada, S. (2016). Teachers' perceptions of implementation of aided AAC to support expressive communication in South African special schools: A pilot investigation. *Augmentative and Alternative Communication*, *32*(4), 282–304. https://doi.org/10.1080/07434618.2016.1246609

Tönsing, K., Van Niekerk, K., Schlünz, G. I., & Wilken, I. (2018). AAC services for multilingual populations: South African service provider perspectives. *Journal of Communication Disorders*, *73*, 62–76. https://doi.org/10.1016/j.jcomdis.2018.04.002

Tönsing, K., van Niekerk, K., Schlünz, G., & Wilken, I. (2019). Multilingualism and augmentative and alternative communication in South Africa – Exploring the views of persons with complex communication needs. *African Journal of Disability*, *8*(0). https://doi.org/10.1017/9781108943024

*Tulsiani, H., Swarup, P., & Rao, P. (2017). Acoustic and language modeling for children's read speech assessment. *National Conference on Communications (NCC)*, 1–6.

*Umesh, S., Sinha, R., & Rama Sanand, D. (2007). Using vocal-tract length normalisation in recognition of children speech. *National Conference on Communications*.

United Nations. (1989). *Convention on the Rights of a Child*.

van Niekerk, K., Dada, S., & Tönsing, K. (2019). Influences on selection of assistive technology for young children in South Africa: Perspectives from rehabilitation professionals. *Disability and Rehabilitation*, *41*(8), 912–925. https://doi.org/10.1080/09638288.2017.1416500

*Vaz, M., Brandl, H., Joublin, F., & Goerick, C. (2009). *Speech imitation with a child's voice: Addressing the correspondence problem*. Proceedings of 13th International Conference on Speech and Computer (SPECOM) (21-25 June 2009), St. Petersburg, Russia.

Visagie, S., Scheffler, E., Seymour, N., & Mji, G. (2020). Assistive technology service delivery in South Africa: Conceptualising a systems approach. *South African Health Review*, *1*, 119–127.

*Von Berg, S., Panorska, A., Uken, D., & Qeadan, F. (2009). DECtalk™ and VeriVox™: Intelligibility, Likeability, and Rate Preference Differences for Four Listener Groups. *Augmentative and Alternative Communication*, *25*(1), 7–18.

Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., Sahidullah, M., Vestman, V., Kinnunen, T., Aik Lee, K., Juvela, L., Alku, P., Peng, Y.-H., Hwang, H.-T., Tsao, Y., Wang, H.-M., Le Maguer, S., Becker, M., Henderson, F., … Ling, Z.-H. (2020). ASVspoof 2019: A large-scale public database of synthetic, converted and replayed speech. *arXiv*. https://doi.org/10.48550/arXiv.1911.01601

Wang, Y., Skerrv-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. (2017). Tacotron: Towards end-to-end speech synthesis. *ArXiv preprint arXiv:1703.10135*. https://doi.org/10.48550/arXiv.1703.10135

Watt, D., Harrison, P., & Cabot-King, L. (2019). Who owns your voice? Linguistic and legal perspectives on the relationship between vocal distinctiveness and the rights of the individual speaker. *The International Journal of Speech, Language and the Law*, *26*(2), 137–180. https://doi.org/10.1558/ijsll.40571

*Watts, O., Yamagishi, J., King, S., & Berkling, K. (2010). Synthesis of child speech with HMM adaptation and voice conversion. *IEEE Transactions on Audio, Speech and Language Processing*, *18*(5), 1005–1016.

*Wendt, O., Hsu, N., Simon, K., Dienhart, A., & Cain, L. (2019). Effects of an iPad-based speech-generating device infused into instruction with the picture exchange communication system for adolescents and young adults with severe autism spectrum disorder. *Behavior Modification*, *43*(6), 898–932.

Wester, M., Wu, Z., & Yamagishi, J. (2015). Human vs machine spoofing detection on wideband and narrowband data. *Paper Presented at Interspeech, Dresden, Germany*, 2047–2051.

Western Cape Language Committee. (2020). *Western Cape Language Policy*. Western Cape Government. https://www.westerncape.gov.za/assets/departments/cultural-affairs-sport/western_cape_language_policy.pdf

World Medical Association (2013) World Medical Association declaration of Helsinki: Ethical principles for medical research involving human subjects. *Clinical Review & Education* 310(20): 2191–2194.

Wormnaes S and Malek YA (2004) Egyptian speech therapists want more knowledge about augmentative and alternative communication. *Augmentative and Alternative Communication* 20: 30–41.

Yamagishi, J., Veaux, C., King, S., & Renals, S. (2012). Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. *Acoustic Science and Technology*, *33*(1).

Yamagishi, J., Watts, O., King, S., & Usabaev, B. (2010). Roles of the Average Voice in Speaker-adaptive HMM-based Speech Synthesis. *Interspeech*.

Yang, H., Oura, K., Wang, H., Gan, Z., & Tokuda, K. (2015). Using speaker adaptive training to realize Mandarin-Tibetan cross-lingual speech synthesis. *Multimedia Tools and Applications: An International Journal*, *74*(22), 9927–9942.

6. Appendices

APPENDIX A: Letter to LSEN principal

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Studies • Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

RE: Research study information

Dear Principal

My name is Camryn Terblanche, I am currently doing my PhD degree in Speech and Language Pathology at the University of Cape Town. As part of the requirement to complete my studies, I am conducting research supervised by Prof Michal Harty and Prof Michelle Pascoe. We have received permission from the Western Cape Education Department to conduct research at several schools in the Western Cape. We would like to ask permission to conduct research at your school.

The title of the study is:
"Feasibility of individualised synthetic speech for children with complex communication needs in three South African languages (South African English, Afrikaans and isiXhosa)."

The aim of this study is to create unique synthetic/computer-made voices for children ages 9;0-13;0 years old who have complex communication needs. This voice will then be placed in an augmentative and alternative communication device (AAC), which generates speech when you click on a button. Children need a home language of South African English, Afrikaans and/or isiXhosa as the synthetic voices will be made in the child's home language and will give them a functional way to communicate, which should improve their academic performance.

If you have a speech and language therapist (SLTs) on site, we will ask them to identify children who may be suitable for the study. Once SLTs identify suitable children, parents will be asked to give written consent for their children to participate before screening may be conducted. Based on the SLT feedback, we will screen a number of children, using the Test of Aided-Communication Symbol Performance, to see if they can use AAC to communicate. In the end, we will only select **three** children for the study (one in each language, from various schools). The children who have the most speech will be given preference. Results from the screening procedure will be individually sent to all parents and used to provide feedback to the teachers on the most appropriate AAC communication methods in the school setting.

Teachers will be offered an optional information/training session related to supporting a child's communication attempts through AAC. The information session will be done at the school and will hopefully assist with student communication and participation in the classroom. The information session will not interfere with their teaching time, will not be longer than an hour, and will be completely optional.

If one of the three children is selected from your school, we would like to make sure our research is relevant. We therefore ask to interview the children's teachers, family members and health professionals. We would like to gather their personal opinions about AAC use in South Africa. We would like to invite them to two interview groups and will ask for their consent before interviewing them. Family members, teachers and health professionals will be able to listen to the final computer-made voices in the second interview group and will be invited to a final feedback session where the project results will be showcased.

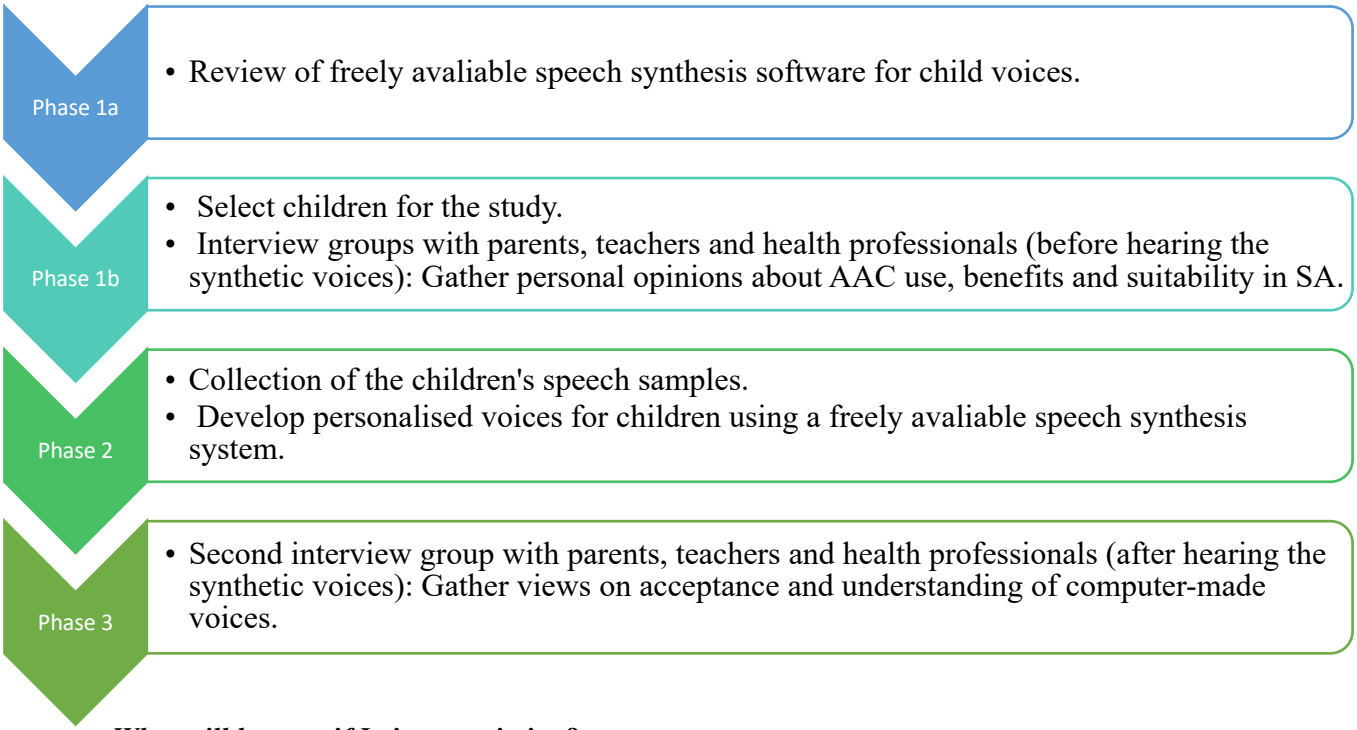
There are no risks to the children or adults participating in this study and personal details will remain confidential. COVID-19 protocols will be followed at all times.

Study information:

What is the importance of the study?

- Creating unique computer-made voices for children has not been done for some South African languages such as isiXhosa, and very little is known about Afrikaans and SAE.
- Describing how to create computer-made voices for children, especially for languages in SA, will be helpful for the SA community and children struggling with communication difficulties.

An overview of the entire study:



What will happen if I give permission?

- There are no risks to the people taking part in this study and personal details will remain confidential.
- We will ensure adequate Covid-19 protocols are followed at all times.

Children:

- Several children will be screened using the Test of Aided-Communication Symbol Performance, to see if they can use AAC to communicate.

- In total, only three children (from various schools) will be asked to participate.
- A speech sample will be collected from the final three children.

Family members/teachers/health professionals:

- Teachers will be offered an optional information session related to supporting a child's communication attempts through AAC, no matter if the child is selected.
- The three children's family members, teachers and health professionals will be asked to participate.
- Teachers and health professionals working at the school will be interviewed twice.

Why and when do we want the children to participate?

- We need the children's speech samples to make the synthetic voices as similar to child's voice as possible.
- The three selected children will have a speech sample collected from them in their school environment in 2022.

Why do we want the family members, teachers, and health professionals to participate?

- We would like to gather the unique views and experiences of family members, teachers and health professionals living with or working with children who have complex communication needs. As they are the child's main communication partners, we would like to take both the communication partner's needs into consideration, along with the child's needs. This will ensure that our research is relevant.
- The groups will be quite small, so as to allow for in-depth discussion. Family members will form one group of approximately 6 people (2 family members for each child selected) and another group will be made up of the health professionals and teachers, which will also be approximately 6 people.

Where, when, and how will the group discussion happen?

- The adults will be asked to participate in two groups. One group will meet in 2022 and the other will meet in 2023.
- Each group could take between 1-2 hours.
- It will take place in a school in Cape Town. If people taking part need to travel to the school, a travel reimbursement will be provided.
- Afrikaans and isiXhosa translators will be present in the groups, so although the researcher will be speaking English, people may speak in the language they feel most comfortable speaking in.

Do I have to take part?

No! You are free to say no. You are free to withdraw permission at any time without having to give us any explanation. In addition, there is a procedure that you can access if you have a complaint.

It is very important to us that you are comfortable with this. If you have any questions, however trivial, please email c.terblanche04@gmail.com.

What will happen after participating in this study?

If the children are selected for the study, they will be provided with a unique computer-made voice that represents who they are, and the teachers will benefit from an optional information session related to

supporting a child's communication attempts through AAC. No one except the researcher will know that your students or teachers took part in this study.

Declaration	**Yes**	**No**
I give institutional permission allowing the researchers to conduct the above mentioned study at …….(institution name).		

Signed: Name_____ Signature_____

Designation: _____

Date __/___/_____

Please feel free to ask the researchers any questions that you may have about this research project.

Ms Camryn Terblanche (Researcher)
Email: c.terblanche04@gmail.com

Prof Michal Harty (Supervisor)
Email: michal.harty@uct.ac.za

Prof Michelle Pascoe (Co- Supervisor)
Email: michelle.pascoe@uct.ac.za

You may contact Professor Marc Blockman, the chairperson of the Faculty of Health Sciences Human Research Ethics Committee on 021 406 6338 if you have any ethical queries or concerns. Thank you for your time and consideration.

Camryn Terblanche

APPENDIX B: Letter to mainstream principal

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Studies • Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

RE: Research study information

Dear Principal

My name is Camryn Terblanche, I am currently doing my PhD degree in Speech and Language Pathology at the University of Cape Town. As part of the requirement to complete my studies, I am conducting research supervised by Prof Michal Harty and Prof Michelle Pascoe. We have received permission from the Western Cape Education Department to conduct research at several schools in the Western Cape. We would like to ask permission to conduct research at your school.

The title of the study is:
"Feasibility of individualised synthetic speech for children with complex communication needs in three South African languages (South African English, Afrikaans and isiXhosa)."

The aim of this study is to create unique synthetic/computer-made voices for children ages 9;0-13;0 years old who have complex communication needs. This voice will then be placed in an augmentative and alternative communication device (AAC), which generates speech when you click on a button. Children who have a home language of South African English, Afrikaans and/or isiXhosa will be selected, as the synthetic voices will be made in each of these languages. This will give the children with complex communication needs a functional way to communicate, which should improve their academic performance.

To make the synthetic speech sound like a typical child, we need to include typically developing children's speech. We would like to ask your permission to audio record a quick session (+-10min) with children aged between 9-13 years old at your school. The children will be asked to describe a picture, and their responses will be audio recorded. Before we begin, we will ask parents to give written consent for their children to participate. Although the research will ultimately focus on the audio recordings, we will also gain some important information about the children's language and narrative skills, which is an important part of literacy development. This information, and what it means for a child's literacy development, will be sent home to the parents, and also outlined to the teachers in an optional information session. The teachers will be offered ways to improve literacy skills in the classroom, and if applicable, students who are struggling will be identified and referred for further assistance. There are no risks to the children participating in this study and personal details will remain confidential. COVID-19 protocols will be followed at all times.

Study information:

What is the importance of the study?

- Creating unique computer-made voices for children has not been done for some South African languages such as isiXhosa, and very little is known about Afrikaans and SAE.
- Describing how to create computer-made voices for children, especially for languages in SA, will be helpful for the SA community and children struggling with communication difficulties.

What will happen if I give permission?

- Several children will be asked to participate in a picture description task, and their responses will be audio recorded.

Why and when do we want the children to participate?

- We need the children's speech samples to make the synthetic voices appear more child-like.
- The speech samples will be collected as soon as possible, before the June-July school holidays in 2022 (and not during the exam period).

Do I have to take part?

No! You are free to say no. You are free to withdraw permission at any time without having to give us any explanation. In addition, there is a procedure that you can access if you have a complaint. It is very important to us that you are comfortable with this. If you have any questions, however trivial, please email c.terblanche04@gmail.com.

What will happen after participating in this study?

If the children participate in the picture description task, we will learn about their language and narrative skills, and the teachers will benefit from an optional information session related to supporting a child's literacy skills in the classroom. No one except the researcher will know that your students or teachers took part in this study.

Declaration	**Yes**	**No**
I give institutional permission allowing the researchers to conduct the above mentioned study at ……(institution name).		

Signed: Name_____ Signature_____

Designation: _____

Date __/___/_____

Please feel free to ask the researchers any questions that you may have about this research project.

Ms Camryn Terblanche (Researcher)
Email: c.terblanche04@gmail.com

Prof Michal Harty (Supervisor)
Email: michal.harty@uct.ac.za

Prof Michelle Pascoe (Co- Supervisor)
Email: michelle.pascoe@uct.ac.za

You may contact Professor Marc Blockman, the chairperson of the Faculty of Health Sciences Human Research Ethics Committee on 021 406 6338 if you have any ethical queries or concerns.

Thank you for your time and consideration.

Camryn Terblanche

APPENDIX C: Letter to SLTs re participant selection

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Studies • Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

Re: Participant selection

Dear SLT

My name is Camryn Terblanche, I am currently doing my PhD degree in Speech and Language Pathology at the University of Cape Town. As part of the requirement to complete my studies, I am conducting research supervised by Prof Michal Harty and Prof Michelle Pascoe. We have received permission from the Western Cape Education Department to conduct research at several schools in the Western Cape and we have also received permission from the principal to conduct research at the school.

The title of the study is:
"Feasibility of individualised synthetic speech for children with complex communication needs in three South African languages (South African English, Afrikaans and isiXhosa)."

The aim of this study is to develop a comprehensive and feasible method for producing the most intelligible and natural-sounding individualised synthetic voices for SA children with complex communication needs (CCN) who speak South African English (SAE), Afrikaans and isiXhosa, by using existing open-source speech synthesis software. Children with CCN may have developmental conditions such as autism spectrum disorder, cerebral palsy or Down syndrome or they may have an acquired disorder, as a result of a traumatic brain injury or stroke.

This study investigates potential open-source software and techniques for voice output personalisation that could be used in a voice output communication aid or AAC device. The study has a primary focus on synthesizing speech in the languages most spoken in the Western Cape: SAE, Afrikaans, and isiXhosa. The study also focuses on gathering the perceptions of teachers, therapists, and family members before and after the development of the individual voices. Thus, if one of your clients is selected, we may ask you to participate in two focus groups (approximately 1 hour each). Further information can be found below in "study information".

I am currently looking for child participants and I was wondering if you may have any potential candidates from your client list? Would you please look at the criteria below and if acceptable to you, consider forwarding the names of potential candidates to me?

Children participants:
1 x South African English child
1 x Afrikaans child
1 x isiXhosa child

Selection criteria include:

- Between the ages of 9;0-13;0 years old, who demonstrate complex communication needs (CCN).
- Either using an augmentative and alternative communication (AAC) device to communicate, or they should be candidates for an AAC device.
- Be able to attend to a task, understand that a symbol/line drawing represents a word or concept, select a picture from a field of four at minimum.
- Have some residual speech abilities.
- Must be able to use direction selection, such as finger pointing or some type of adapted pointer to indicate their choices.

Children who are still functioning at a concrete object level, cannot recognise pictures as a result of visual or cognitive impairment, have severely limited fine-motor skills, severe auditory processing problems or children who are completely non-vocal, will not be included in the study.

If you have children that meet the selection criteria on your caseload, we will ask the parents to give written consent for their children to participate before screening may be conducted. Based on your feedback, we will screen a number of children, using the Test of Aided-Communication Symbol Performance. In the end, we will only select **three** children for the study (one in each language, from various schools). The children with the most residual and intelligible speech will be given preference. Results from the screening procedure will be individually sent to all parents and used to provide feedback to the teachers on the most appropriate AAC communication methods in the school setting. Family members, teachers and health professionals will be able to listen to the final synthesised voices in the second interview group and will be invited to a final feedback session where the project results will be showcased.

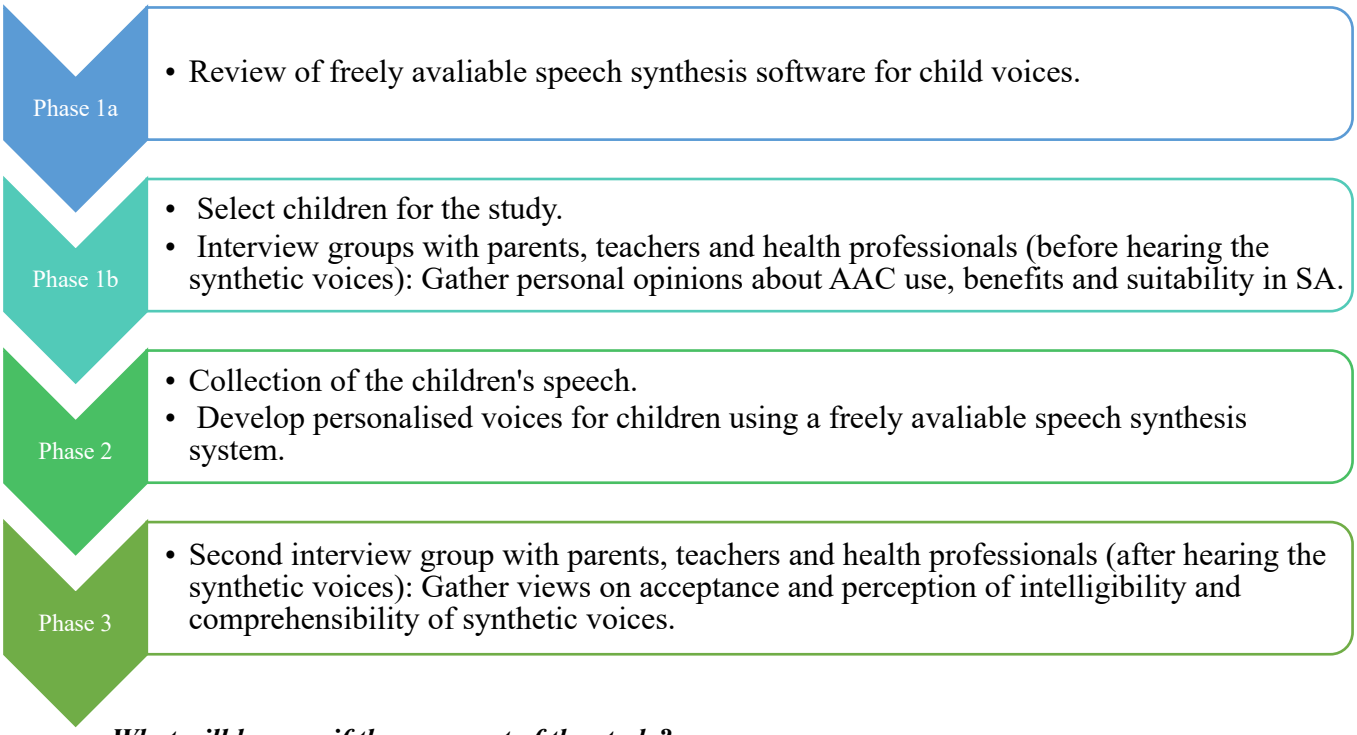
There are no risks to the children or adults participating in this study and personal details will remain confidential. COVID-19 protocols will be followed at all times.

Study information:

What is the importance of the study?

- Creating synthetic voices for children has yet to be done for South African languages such as isiXhosa, and very little is known about Afrikaans and SAE.
- Community perceptions of the speech output will provide a unique perspective for AAC research. This study will form the basis for further research surrounding multilingual speech synthesis systems for children and adults.
- This research could form the foundation for the design and creation of a relatively simple voice output communication aid (VOCA) app that could be used by speech and language therapists who are looking to find unique voices for their patient's systems.

An overview of the entire study:



What will happen if they are part of the study?

- There are no risks to the people taking part in this study and personal details will remain confidential.
- We will ensure adequate Covid-19 protocols are followed at all times.

Children:

- Several children will be screened using the Test of Aided-Communication Symbol Performance, to see if they can use AAC to communicate.
- In total, only three children (from various schools) will be asked to participate.
- A speech sample will be collected from the final three children.

Family members/teachers/health professionals:

- Teachers will be offered an optional information session related to supporting a child's communication attempts through AAC, no matter if the child is selected.
- The three children's family members, teachers and health professionals will be asked to participate.
- Teachers and health professionals working at the school will be interviewed twice.

Why and when do we want the children to participate?

- We need the children's speech samples to make the synthetic voices as similar to child's voice as possible.
- The three selected children will have a speech sample collected from them in their school environment in 2022.

Why do we want the teachers and health professionals to participate?

- We would like to gather the unique views and experiences of family members, teachers and health professionals living with or working with children who have complex communication needs. As they are the child's main communication partners, we would like to take both the communication

partner's needs into consideration, along with the child's needs. This will ensure that our research is relevant.

- The groups will be quite small, so as to allow for in-depth discussion. Family members will form one group of approximately 6 people (2 family members for each child selected) and another group will be made up of the health professionals and teachers, which will also be approximately 6 people.

Where, when, and how will the group discussion happen?

- The adults will be asked to participate in two groups. One group will meet in 2022 and the other will meet in 2023.
- Each group could take between 1-2 hours.
- It will take place in a school in Cape Town. If people taking part need to travel to the school, a travel reimbursement will be provided.
- Afrikaans and isiXhosa translators will be present in the groups, so although the researcher will be speaking English, people may speak in the language they feel most comfortable speaking in.

Do I have to take part?

No! You are free to say no. You are free to withdraw at any time without having to give us any explanation. In addition, there is a procedure that you can access if you have a complaint. It is very important to us that you are comfortable with this. If you have any questions, however trivial, please email c.terblanche04@gmail.com.

What will happen to me or the children after participating in this study?

There are anticipated journal articles and conference presentations, and your experience would be greatly appreciated. No one except the researcher will know that you took part in this study. If the children are selected for the study, they will be provided with a personalised synthetic voice that represents who they are.

Please feel free to ask the researchers any questions that you may have about this research project.

Ms Camryn Terblanche (Researcher)
Email: c.terblanche04@gmail.com

Prof Michal Harty (Supervisor)
Email: michal.harty@uct.ac.za

Prof Michelle Pascoe (Co- Supervisor)
Email: michelle.pascoe@uct.ac.za

You may contact Professor Marc Blockman, the chairperson of the Faculty of Health Sciences Human Research Ethics Committee on 021 406 6338 if you have any ethical queries or concerns.

Thank you for your time and consideration.

Camryn Terblanche

APPENDIX D: Letter and consent form to SLT re study information

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Studies • Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

RE: Research study information

Dear SLT

My name is Camryn Terblanche, I am currently doing my PhD degree in Speech and Language Pathology at the University of Cape Town. As part of the requirement to complete my studies, I am conducting research supervised by Prof Michal Harty and Prof Michelle Pascoe.

The title of the study is:
"Feasibility of individualised synthetic speech for children with complex communication needs in three South African languages (South African English, Afrikaans and isiXhosa)."

The aim of this study is to develop a comprehensive and feasible method for producing the most intelligible and natural-sounding individualised synthetic voices for SA children with CCN who speak South African English (SAE), Afrikaans and isiXhosa, by using existing open-source speech synthesis software. Children with CCN may have developmental conditions such as autism spectrum disorder (ASD), cerebral palsy (CP) or Down syndrome or they may have an acquired disorder, as a result of a traumatic brain injury (TBI) or stroke. The study also focuses on gathering the perceptions of teachers, therapists, and parents before and after the synthetic voice development.

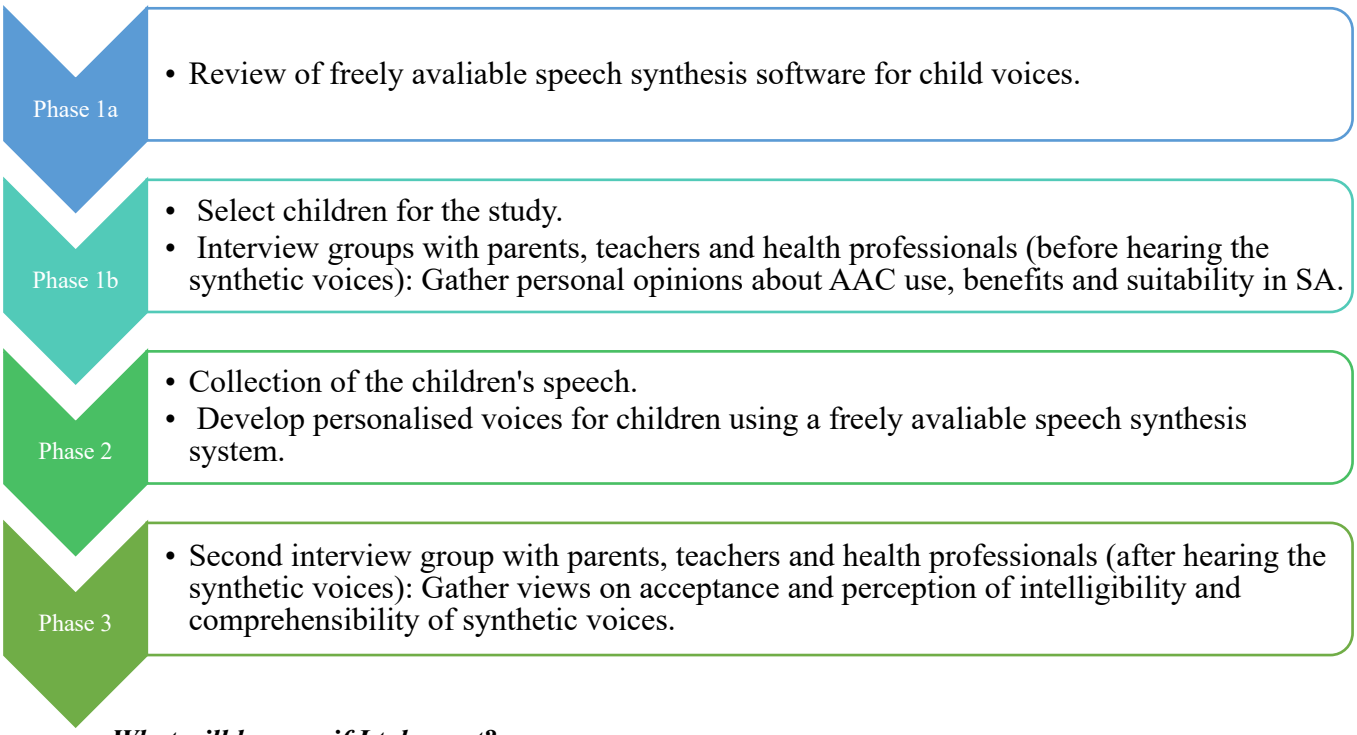
I would like to invite you to volunteer to participate in two focus groups. The focus groups will be recorded and later transcribed. The information and identities will be protected through the use of a coding system. All documents will be stored in a private location and password protected. The study is voluntary, and any person can decide to pull out without any penalties or judgement.

Study information:

What is the importance of the study?

- Creating synthetic voices for children has yet to be done for South African languages such as isiXhosa, and very little is known about Afrikaans and SAE.
- Community perceptions of the speech output will provide a unique perspective for AAC research. This study will form the basis for further research surrounding multilingual speech synthesis systems for children and adults.
- This research could form the foundation for the design and creation of a relatively simple voice output communication aid (VOCA) app that could be used by speech and language therapists who are looking to find unique voices for their patient's systems.

An overview of the entire study:



What will happen if I take part?

If you decide that you would like to get involved, we will ask you to do two things:

- Firstly, we will ask you to fill in our consent form. This is a form that says that you agree to take part in the study.
- Secondly, we will ask you to participate in two focus groups (approximately 1 hour each).
    - The first focus group will occur before hearing the computer-made voices. We will ask you your opinions related to communication devices which have a voice output capability.
    - The second focus group will occur after you have heard the computer-made voices. This will involve listening to the quality of those voices and deciding if they would be useful for children at home and at school.

Why do we want the teachers and health professionals to participate?

- We would like to gather the unique views and experiences of family members, teachers and health professionals living with or working with children who have complex communication needs. As they are the child's main communication partners, we would like to take both the communication partner's needs into consideration, along with the child's needs. This will ensure that our research is relevant.
- The groups will be quite small, so as to allow for in-depth discussion. Family members will form one group of approximately 6 people (2 family members for each child selected) and another group will be made up of the health professionals and teachers, which will also be approximately 6 people.

Where, when, and how will the group discussion happen?

- The adults will be asked to participate in two groups. One group will meet over the next few weeks in 2022 and the other will meet in 2023.
- Each group could take between 1-2 hours.

- It will take place in a school in Cape Town. If people taking part need to travel to the school, a travel reimbursement will be provided.

Do I have to take part?
No! You are free to say no. You are free to withdraw at any time without having to give us any explanation. In addition, there is a procedure that you can access if you have a complaint. It is very important to us that you are comfortable with this. If you have any questions, however trivial, please email c.terblanche04@gmail.com.

What will happen to me or the children after participating in this study?
There are anticipated journal articles and conference presentations, and your experience would be greatly appreciated. No one except the researcher will know that you took part in this study. If the children are selected for the study, they will be provided with a personalised synthetic voice that represents who they are.

Please feel free to ask the researchers any questions that you may have about this research project.

Ms Camryn Terblanche (Researcher)
Email: c.terblanche04@gmail.com

Prof Michal Harty (Supervisor)
Email: michal.harty@uct.ac.za

Prof Michelle Pascoe (Co- Supervisor)
Email: michelle.pascoe@uct.ac.za

You may contact Professor Marc Blockman, the chairperson of the Faculty of Health Sciences Human Research Ethics Committee on 021 406 6338 if you have any ethical queries or concerns.

Thank you for your time and consideration.

Camryn Terblanche

NB: please complete and sign consent/assent form on next page

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Studies • Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

Consent form

I, _____, hereby give permission to participate in the above research study. I completely understand the purpose of this study and what it entails has been thoroughly explained to me.

I understand that:
This study will require me to participate in two focus groups. The first will comment on voice output communication aids in South Africa. The second will involve reviewing the quality of synthesised voices and determining their applicability in different settings. I understand that my participation is voluntary and that I may choose to withdraw at any time without giving a reason. Any information I give will be kept strictly confidential and no names will be used to identify me in this study without my approval.

Declaration	**Yes**	**No**
I have read through the information sheet and understand it's content		
I understand that my consent is required		
I understand that participation is voluntary, and I can withdraw from the study without any consequences		
I understand that I will not be personally identified should this research study be published		
I consent to participating in this research study of my own free will		
I understand that I will not be paid for my participation		
I consent to having the focus group discussion recorded		

Participant signature: _____
Date: _____/_____/_____

Researcher signature: _____
Date: _____/_____/_____

APPENDIX E: Letter and consent form to parents/guardian of children with CCN re screening (English example)

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Studies • Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

RE: Research study information

Dear Parent/Guardian

My name is Camryn Terblanche, I am currently doing my PhD degree in Speech and Language Pathology at the University of Cape Town. As part of my studies, I am doing research with Prof Michal Harty and Prof Michelle Pascoe.

The title of the study is:
"Feasibility of individualised synthetic speech for children with complex communication needs in three South African languages (South African English, Afrikaans and isiXhosa)."

The aim of this study is to create unique computer-made voices for children who have trouble talking, using free software from the internet. The children need to speak South African English (SAE), Afrikaans and/or isiXhosa. The children who have trouble talking may have developmental conditions such as autism spectrum disorder, cerebral palsy or Down syndrome or it may have come about later, because of a traumatic brain injury or stroke. The study also focuses on gathering the thoughts of teachers, family members, and therapists before and after the computer-made voices have been created.

Your child has been identified by their speech and language therapist as someone who may be able to benefit from the study. I would like to invite you to volunteer your child for **screening**.

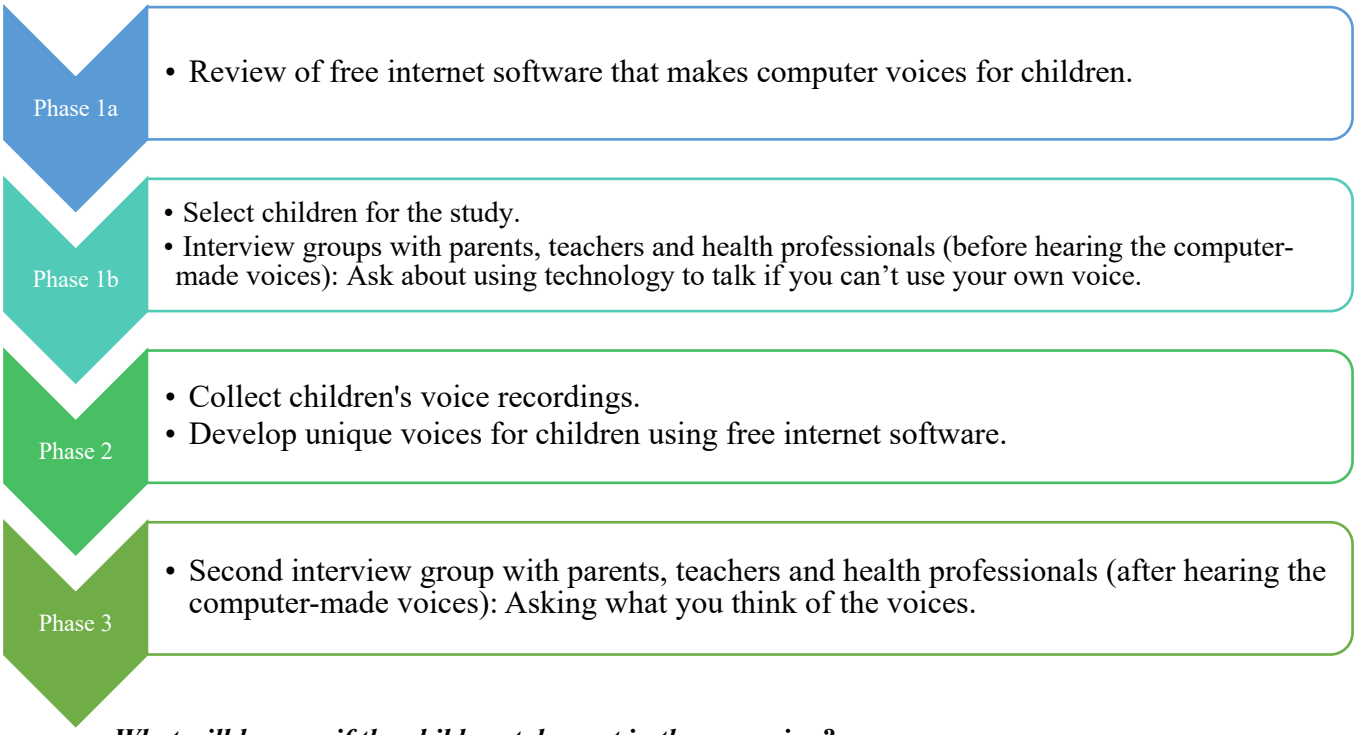
If you agree to allow your child to be screened, it **DOES NOT necessarily mean that they will receive a computer-made voice**. We are screening a lot of children but are just looking for THREE children. The three children that we will choose in the end will be able to say the most, and their speech will be as clear as possible so that it can be voice recorded.

Study information:

What is the importance of the study?

- Creating unique computer-made voices for children has not been done for some South African languages such as isiXhosa, and very little is known about Afrikaans and SA English.
- Describing how to make computer-made voices for children, especially for languages in SA, will be helpful for both the SA community and children who struggle to talk.

An overview of the entire study:



What will happen if the children take part in the screening?

- The children will be tested on their ability to choose the right picture from a group of pictures. This test will tell us if your child would be able to use a different form of communication, to help their talking.
- The children will be screened/ tested at their school by the researcher, who is a speech and language therapist. The school speech therapist and/or teacher can be present.
- It will happen in the next couple months of 2022.
- It will take between 10-20 minutes to finish.
- Remember, if your child is screened, it **DOES NOT necessarily mean that they will receive a computer-made voice**. We are just looking for three children. The three children that we will choose in the end will be able to say the most, and their speech will be as clear as possible so that it can be recorded.

What will happen if my child is screened but NOT selected?

- The teachers will be offered training on the best way to help your child communicate at school.
- The screening results will be sent home with the children.

What will happen if my child is screened and then selected as one of the three children?

- The teachers will be offered training on the best way to help your child communicate at school.
- The screening results will be sent home with the children.
- A voice recording of your child's voice will be collected at their school. If you like, you are welcome to be present while we make the recordings.

- If your child is selected, their family members will also be asked to participate, along with their teachers and therapists.
- We will ask you to take part in two interview groups (about 1 hour each). One group will happen in 2022 and the other will happen in 2023.

- The first group will happen before you hear the computer-made voices. We will ask you what you think about using technology to talk if you can't use your own voice.
- The second group will happen after you have heard the computer-made voices. Family members, health professionals and teachers will listen to all the voices made and say if they think they could be used at school. We will also ask you to fill in our consent form. This is a form that says that you agree to take part in the study.
- The group interviews will take place in a school in Cape Town. If you need to travel to the school, you will get the travel money back.
- Afrikaans and isiXhosa translators will be in the groups, so although I will be speaking English, you may speak in the language you feel most comfortable speaking in.
- We will ensure Covid-19 protocols are always followed.

Why and when do we want the children to participate?

- We need the children's voice recordings to make computer voices that sounds like the child's voice.
- The three final children will have their voices recorded in their school in 2022.

Why do we want the family members, teachers and health professionals to participate?

- We would like to gather the unique views of family members, teachers and therapists living with or working with children who have a difficulty talking.
- As they talk to the child the most, we would like to take both their needs into consideration, along with the child's needs. This will ensure that our research is relevant.
- The interview groups will be quite small. Family members will form one group of about 6 people (2 family members for each child selected) and another group will be made up of the therapists and teachers, which will also be about 6 people.

Do I have to take part?

No! You are free to say no. You are free to stop taking part at any time without having to give us a reason. It is very important to us that you are comfortable with this. If you do not want to take part, it will not affect your child's therapy or education in any way. This information will not be shared with the school. If you have any questions, please email c.terblanche04@gmail.com.

What will happen to me or the children after participating in this study?

The children will benefit from the screening because we will learn if they are ready to use a communication board. After screening, you will be provided with the results and the teachers will be offered training on the best way to help your child communicate at school. The data collected will remain anonymous.

Please feel free to ask the researchers any questions that you may have about this research project.

Ms Camryn Terblanche (Researcher)
Email: c.terblanche04@gmail.com

Prof Michal Harty (Supervisor)
Email: michal.harty@uct.ac.za

Prof Michelle Pascoe (Co- Supervisor)
Email: michelle.pascoe@uct.ac.za

You may contact Professor Marc Blockman, the chairperson of the Faculty of Health Sciences Human Research Ethics Committee on 021 406 6338 if you have any ethical queries or concerns.

Thank you for your time and consideration.

Camryn Terblanche

NB: please complete and sign consent form on next page

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Studies • Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

Consent form

I, _____, hereby give permission for my child to be screened. I completely understand the purpose of this study and what it entails has been thoroughly explained to me.

I understand that:
My child will be screened/tested to determine their communication level. This is done to see if they would be able to use alternative communication devices appropriately. Being screened does NOT mean that they will receive a computer-made voice.

Declaration	**Yes**	**No**
I have read through the information sheet and understand its content		
I understand that my consent is required		
I understand that participation is voluntary, and I can withdraw at any time		
I understand that neither I nor my child will be personally identified should this research study be published		
I give permission for my child to be screened		
I understand that we will not be paid for the screening		
I understand that this is just a test and does not mean that my child will receive a computer-made voice		

Participant signature: _____
Date: _____/_____/_____

Researcher signature: _____
Date: _____/_____/_____

Children Assent

I will ask you some questions, is that okay?	 YES  NOT SURE  NO

APPENDIX F: Letter and consent form for parents/guardians of children with CCN (English example)

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Studies • Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

RE: [Anonymous] child

Dear Parent/Guardian

My name is Camryn Terblanche, I am currently doing my PhD degree in Speech and Language Pathology at the University of Cape Town. As part of my studies, I am doing research with Prof Michal Harty and Prof Michelle Pascoe.

The title of the study is:
"Feasibility of individualised synthetic speech for children with complex communication needs in three South African languages (South African English, Afrikaans and isiXhosa)."

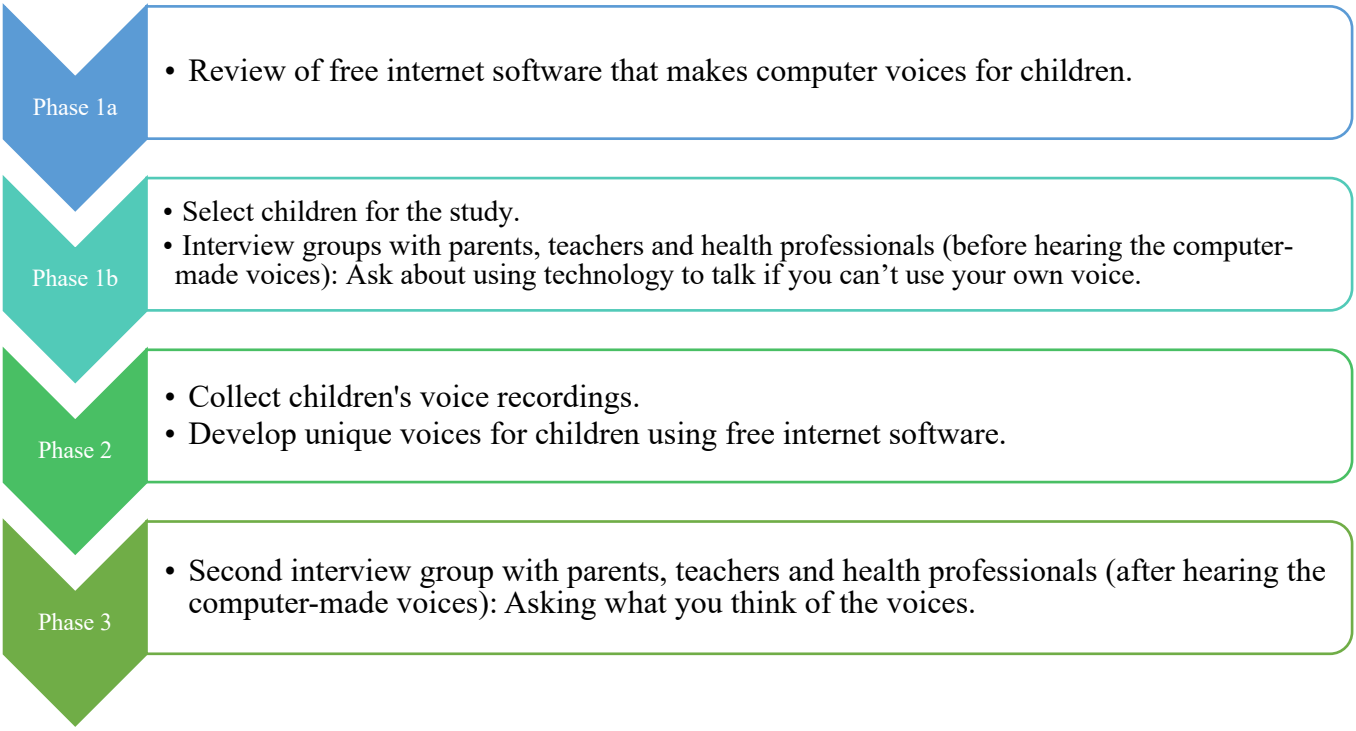
The aim of this study is to create unique computer-made voices for children who have trouble talking, using free software from the internet. The children need to speak South African English (SAE), Afrikaans and/or isiXhosa. The children who have trouble talking may have developmental conditions such as autism spectrum disorder, cerebral palsy or Down syndrome or it may have come about later, because of a traumatic brain injury or stroke. The study also focuses on gathering the thoughts of teachers, family members, and therapists before and after the computer-made voices have been created.

After screening [anonymous child], we have selected him to be part of the study. A computer-made voice will be created for them. I would like to ask permission for your child, as well as to invite you to take part in two interview groups yourself. We ask that two family members per child take part (parents/guardians, grandparents, siblings, aunts, uncles, close family friends etc.) The groups will be voice recorded and responses will be written down. No-one will know that you took part. The study is voluntary, and any person can decide to pull out without judgement or anything happening to them.

Study information:
What is the importance of the study?

- Creating unique computer-made voices for children has not been done for some South African languages such as isiXhosa, and very little is known about Afrikaans and SA English.
- Describing how to make computer-made voices for children, especially for languages in SA, will be helpful for both the SA community and children who struggle to talk.

An overview of the entire study:



What will happen if they take part in the study?

- There are no risks to the people taking part in this study and personal details will remain confidential.
- We will ensure adequate Covid-19 protocols are followed at all times.

Children:

- A voice recording of your child's speech will be collected. Your child will look at pictures and tell us what they can see.

Parents/Teachers/health professionals:

- The children's parents, teachers and health professionals will be asked to participate in two interview groups.

Why and when do we want the children to participate?

- We need the children's voice recordings to make computer voices that sounds like the child's voice.
- The three final children will have their voices recorded in their school in 2022.

Why and when do we want the family members, teachers and health professionals to participate?

- The child's family members (2 per family) will be asked to take part in two interview groups. One group will happen in 2022 and the other will happen in 2023.
- The first group will happen before you hear the computer-made voices. We will ask you what you think about using technology to talk if you can't use your own voice.
- After we have made the children's voices, the second group will happen after you have heard the computer-made voices. The parents, teachers and health professionals will listen to the voices and say if they think they could be used at school.

- The interview groups will be quite small. Family members will form one group of about 6 people (2 family members for each child selected) and another group will be made up of the therapists and teachers, which will also be about 6 people.
- Each group could be between 1-2 hours long.
- It will take place in a school in Cape Town. If you need to travel to the school, you will get the travel money back.
- Afrikaans and isiXhosa translators will be in the groups, so although the researcher will be speaking English, you may speak in the language you feel most comfortable speaking in.
- We will ensure Covid-19 protocols are always followed.

Do I have to take part?
No! You are free to say no. You are free to stop taking part at any time without having to give us a reason. It is very important to us that you are comfortable with this and that you feel confident filling in the forms. If you do not want to take part, it will not affect your child's therapy or education in any way. This information will not be shared with the school. If you have any questions, please email c.terblanche04@gmail.com.

What will happen to me or the children after participating in this study?
Your child will be given a unique computer-made voice that represents who they are at the end of the study. We plan to write about what is spoken about in the interview groups, but we will not mention your name. No one except the researcher will know that either of you took part in this study.

Please feel free to ask the researchers any questions that you may have about this research project.

Ms Camryn Terblanche (Researcher)
Email: c.terblanche04@gmail.com

Prof Michal Harty (Supervisor)
Email: michal.harty@uct.ac.za

Prof Michelle Pascoe (Co- Supervisor)
Email: michelle.pascoe@uct.ac.za

You may contact Professor Marc Blockman, the chairperson of the Faculty of Health Sciences Human Research Ethics Committee on 021 406 6338 if you have any ethical queries or concerns.

Thank you for your time and consideration.

Camryn Terblanche

NB: please complete and sign consent form on next page

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Studies • Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

Consent form

I, _____, hereby give permission to participate in the above research study. I completely understand the purpose of this study and what it entails has been thoroughly explained to me.

I understand that:
This study will require my family participate in two interview groups. The first will look at AAC use in South Africa. The second will involve listening to the quality of the computer-made voices. I understand that my child will also be a part of the study and I'm providing permission for that, including a speech assessment and the development of a computer-made voice. I understand that our participation is voluntary and that I may choose to stop participating at any time without giving a reason. Any information I give will be kept strictly confidential and that no names will be used to identify me in this study without my approval.

Declaration	**Yes**	**No**
I have read through the information sheet and understand its content		
I understand that my consent is needed		
I understand that both my and my child's participation is voluntary, and either can stop participating from the study without any consequences		
I understand that neither I nor my child will be personally identified should this research study be published		
I consent to participating in this research study and give permission for my child to participate		
I consent to having my child's speech recorded		
I consent to having the group discussion audio recorded		
I understand that we will not be paid for our participation		
I understand the speech samples may be used in future research		

Participant signature: _____
Date: _____/_____/_____

Researcher signature: _____
Date: _____/_____/_____
Date: _____/_____/_____

Children Assent

I will ask you some questions, is that okay?	 YES NOT SURE NO
Is it okay if I record your voice?	 YES NOT SURE NO

APPENDIX G: Letter and consent form to parents/guardians of children in mainstream school re screening

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Studies • Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

RE: Research study information

Dear Parent/Guardian

My name is Camryn Terblanche, I am currently doing my PhD degree in Speech and Language Pathology at the University of Cape Town. As part of my studies, I am doing research with Prof Michal Harty and Prof Michelle Pascoe.

The title of the study is:
"Feasibility of individualised synthetic speech for children with complex communication needs in three South African languages (South African English, Afrikaans and isiXhosa)."

The aim of this study is to create unique computer-made voices for children who have trouble talking, using free software from the internet. The voices will be made for children who speak South African English (SAE), Afrikaans and/or isiXhosa. The children who have trouble talking may have developmental conditions such as autism spectrum disorder, cerebral palsy or Down syndrome or it may have come about later, because of a traumatic brain injury or stroke.

To make the computer-made voices sound natural, we need to include typically developing children's speech. We would like to ask your permission to voice record a quick session with your child (about 10min). The children will be asked to describe a picture, and their answers will be voice recorded. Although my research will focus on the voice recordings, we will also learn a lot about the children's story telling skills, which is an important skill for school success. Feedback will be sent home with the child and the teachers will be given an optional information session. Teachers will be offered ways to improve learning in the classroom, and if needed, students who are struggling will be referred for help with reading. There are no risks to the children participating in this study and personal details will remain confidential. COVID-19 protocols will be followed at all times.

What is the importance of the study?

- Creating unique computer-made voices for children has not been done for some South African languages such as isiXhosa, and very little is known about Afrikaans and SA English.
- Describing how to make computer-made voices for children, especially for languages in SA, will be helpful for both the SA community and children who struggle to talk.

What will happen if I give permission?

- About 30 children (including your child) will be asked to describe a picture, and whatever they say will be voice recorded.

Why and when do we want the children to participate?

- We need the children's voice recordings to make the computer-made voices appear more like a child.
- The recordings will be collected as soon as possible, before the June-July school holidays in 2022 (and not during the exam period).

Do I have to take part?

No! You are free to say no. You are free to stop taking part at any time without having to give us a reason. It is very important to us that you are comfortable with this. If you do not want to take part, it will not affect your child's education in any way. If you have any questions, please email c.terblanche04@gmail.com.

What will happen after participating in this study?

If the children participate in the picture description task, we will learn about their story telling skills. Afterwards, you will be provided with the results and the teachers will be offered training on the best way to help better your child's learning at school. The data collected will remain anonymous.

Please feel free to ask the researchers any questions that you may have about this research project.

Ms Camryn Terblanche (Researcher)
Email: c.terblanche04@gmail.com

Prof Michal Harty (Supervisor)
Email: michal.harty@uct.ac.za

Prof Michelle Pascoe (Co- Supervisor)
Email: michelle.pascoe@uct.ac.za

You may contact Professor Marc Blockman, the chairperson of the Faculty of Health Sciences Human Research Ethics Committee on 021 406 6338 if you have any ethical queries or concerns.
Thank you for your time and consideration.

Camryn Terblanche
NB: please complete and sign consent form on next page

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Studies • Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

Consent form

I, _____, hereby give permission for my child to be screened. I completely understand the purpose of this study and what it entails has been thoroughly explained to me.

I understand that:
My child will be asked to describe a picture, and their answers will be voice recorded. What they say will help us learn about their story-telling skills, and will help the researcher make computer-made voices.

Declaration	**Yes**	**No**
I have read through the information sheet and understand its content		
I understand that my consent is required		
I understand that participation is voluntary, and I can withdraw at any time		
I understand that my child will not be personally identified should this research study be published		
I give permission for my child to be screened		
I give permission for my child's voice to be audio recorded		
I understand that we will not be paid for the screening		

Participant signature: _____
Date: _____/_____/_____

Researcher signature: _____
Date: _____/_____/_____

APPENDIX H: Letter and consent form for parents/guardians of children in mainstream school who were selected

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Studies • Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

RE: Research study information

Dear Parent/Guardian

My name is Camryn Terblanche, I am currently doing my PhD degree in Speech and Language Pathology at the University of Cape Town. As part of my studies, I am doing research with Prof Michal Harty and Prof Michelle Pascoe.

The title of the study is:
"Feasibility of individualised synthetic speech for children with complex communication needs in three South African languages (South African English, Afrikaans and isiXhosa)."

The aim of this study is to create unique computer-made voices for children who have trouble talking, using free software from the internet. The voices will be made for children who speak South African English (SAE), Afrikaans and/or isiXhosa. The children who have trouble talking may have developmental conditions such as autism spectrum disorder, cerebral palsy or Down syndrome or it may have come about later, because of a traumatic brain injury or stroke.

To make the computer-made voices sound natural, we need to include typically developing children's speech. Recently, I saw [anonymous child] at school to screen their narrative and literacy skills. [Anonymous child] did an excellent job and was able to read in [language] at a level above all the other children. Due to this, I am asking your child, and two others, to read for me for a longer time (approximately two hours, and not necessarily in one session). We would like to ask your permission to voice record your child while they are reading out loud. There are no risks to the children participating in this study and personal details will remain confidential. COVID-19 protocols will be followed at all times.

What is the importance of the study?

- Creating unique computer-made voices for children has not been done for some South African languages such as isiXhosa, and very little is known about Afrikaans and SA English.
- Describing how to make computer-made voices for children, especially for languages in SA, will be helpful for both the SA community and children who struggle to talk.

What will happen if I give permission?

- 3 children (including your child) will be asked to read a storybook in their home language, and their reading will be voice recorded.

Why and when do we want the children to participate?

- We need the children's voice recordings to make the computer-made voices appear more like a child.
- The recordings will be collected after school, so as not to interrupt his school time.
- The recordings will be collected as soon as possible, before the June-July school holidays in 2022 (and not during the exam period).

Do I have to take part?

No! You are free to say no. You are free to stop taking part at any time without having to give us a reason. It is very important to us that you are comfortable with this. If you do not want to take part, it will not affect your child's education in any way. If you have any questions, please email c.terblanche04@gmail.com.

What will happen after participating in this study?

The researcher will use your child's audio recordings to create computer-made voices for children with special needs. These computer-made voices will be used to help the children with special needs communicate at school. The speech data collected will remain anonymous.

Please feel free to ask the researchers any questions that you may have about this research project.

Ms Camryn Terblanche (Researcher)
Email: c.terblanche04@gmail.com

Prof Michal Harty (Supervisor)
Email: michal.harty@uct.ac.za

Prof Michelle Pascoe (Co- Supervisor)
Email: michelle.pascoe@uct.ac.za

You may contact Professor Marc Blockman, the chairperson of the Faculty of Health Sciences Human Research Ethics Committee on 021 406 6338 if you have any ethical queries or concerns.

Thank you for your time and consideration.

Camryn Terblanche
NB: please complete and sign consent form on next page

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Studies • Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

Consent form

I, _____, hereby give permission for my child's reading to be audio recorded. I completely understand the purpose of this study and what it entails has been thoroughly explained to me.

I understand that:
My child will be asked to read a story, and their reading will be voice recorded. More than one reading session will be necessary. My child's voice recordings will be used to make computer-made voices for children with special needs.

Declaration	**Yes**	**No**
I have read through the information sheet and understand its content		
I understand that my consent is required		
I understand that participation is voluntary, and that my child can withdraw at any time		
I understand that my child will not be personally identified should this research study be published		
I give permission for my child's voice to be audio recorded		

Participant signature: _____
Date: _____/_____/_____

Researcher signature: _____
Date: _____/_____/_____

APPENDIX I: Letter and consent form to teacher

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Studies • Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

RE: Research study information

Dear Teacher

My name is Camryn Terblanche, I am currently doing my PhD degree in Speech and Language Pathology at the University of Cape Town. As part of the requirement to complete my studies, I am conducting research supervised by Prof Michal Harty and Prof Michelle Pascoe.

The title of the study is:
"Feasibility of individualised synthetic speech for children with complex communication needs in three South African languages (South African English, Afrikaans and isiXhosa)."

The aim of this study is to create unique synthesised/computer-made voices for children who have trouble communicating, using free software from the internet. The children need a language background of South African English (SAE), Afrikaans and/or isiXhosa. The children who have difficulty communicating, or present with complex communication needs (CCN) may have developmental conditions such as autism spectrum disorder (ASD), cerebral palsy (CP) or Down syndrome or it may have come about later, as a result of a traumatic brain injury (TBI) or stroke. The study also focuses on gathering the opinions of teachers, parents/legal guardians, and therapists before and after the synthesised voices have been created.

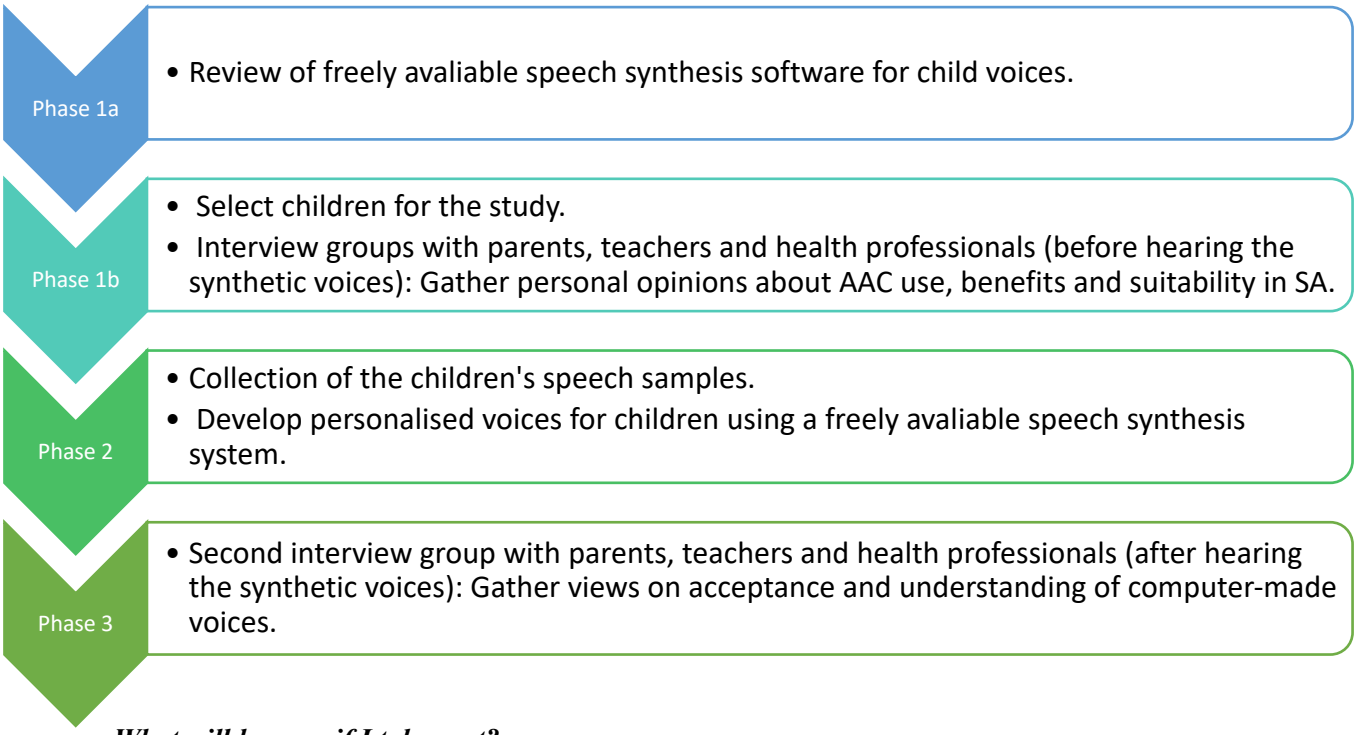
I would like to invite you to volunteer to participate in two focus groups. The focus groups will be recorded and later transcribed. The information and identities will be protected. All documents will be stored in a private location and password protected. The study is voluntary, and any person can decide to pull out without any penalties or judgement.

Study information:

What is the importance of the study?

- Creating unique computer-made voices for children has not been done for some South African languages such as isiXhosa, and very little is known about Afrikaans and SAE.
- Describing how to create computer-made voices for children, especially for languages in SA, will be helpful for the SA community and children struggling with communication difficulties.

An overview of the entire study:



What will happen if I take part?

If you decide that you would like to get involved, we will ask you to do two things:

- Firstly, we will ask you to fill in our consent form. This is a form that says that you agree to take part in the study.
- Secondly, we will ask you to participate in two focus groups (approximately 1 hour each).
    - The first focus group will occur before hearing the computer-made voices. We will ask you your opinions related to communication devices which have a voice output capability.
    - The second focus group will occur after you have heard the computer-made voices. This will involve listening to the quality of those voices and deciding if they would be useful for children at home and at school.

Why do we want the family members, teachers, and health professionals to participate?

- We would like to gather the unique views and experiences of family members, teachers and health professionals living with or working with children who have complex communication needs. As they are the child's main communication partners, we would like to take both the communication partner's needs into consideration, along with the child's needs. This will ensure that our research is relevant.
- The groups will be quite small, so as to allow for in-depth discussion. Family members will form one group of approximately 6 people (2 family members for each child selected) and another group will be made up of the health professionals and teachers, which will also be approximately 6 people.

Where, when, and how will the group discussion happen?

- The adults will be asked to participate in two groups. One group will meet over the next few weeks in 2022 and the other will meet in 2023.
- Each group could take between 1-2 hours.

- It will take place in a school in Cape Town. If people taking part need to travel to the school, a travel reimbursement will be provided.

Do I have to take part?
No! You are free to say no. You are free to withdraw permission at any time without having to give us any explanation. In addition, there is a procedure that you can access if you have a complaint. It is very important to us that you are comfortable with this. If you have any questions, however trivial, please email c.terblanche04@gmail.com.

What will happen after participating in this study?
The children selected for the study will be provided with a unique computer-made voice that represents who they are, and the teachers will benefit from an optional information session related to supporting a child's communication attempts through AAC. No one except the researcher will know that your students or teachers took part in this study.

Please feel free to ask the researchers any questions that you may have about this research project.

Ms Camryn Terblanche (Researcher)
Email: c.terblanche04@gmail.com

Prof Michal Harty (Supervisor)
Email: michal.harty@uct.ac.za

Prof Michelle Pascoe (Co- Supervisor)
Email: michelle.pascoe@uct.ac.za

You may contact Professor Marc Blockman, the chairperson of the Faculty of Health Sciences Human Research Ethics Committee on 021 406 6338 if you have any ethical queries or concerns. Thank you for your time and consideration.

Camryn Terblanche
NB: please complete and sign consent/assent form on next page

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
HEALTH SCIENCES

DHRS
Department of Health and Rehabilitation Sciences
FACULTY OF HEALTH SCIENCES

Divisions of Communication Sciences & Disorders • Disability Studies • Nursing & Midwifery • Occupational Therapy • Physiotherapy

F45 Old Main Building, Groote Schuur Hospital
Observatory, Cape Town, South Africa, 7925
Telephone: +27 (0) 21 406 6401
Website: www.dhrs.uct.ac.za

Consent form

I, _____, hereby give permission to participate in the above research study. I completely understand the purpose of this study and what it entails has been thoroughly explained to me.

I understand that:
This study will require me to participate in two focus groups. The first will comment on voice output communication aids in South Africa. The second will involve reviewing the quality of synthesised voices and determining their applicability in different settings. I understand that my participation is voluntary and that I may choose to withdraw at any time without giving a reason. Any information I give will be kept strictly confidential and that no names will be used to identify me in this study without my approval.

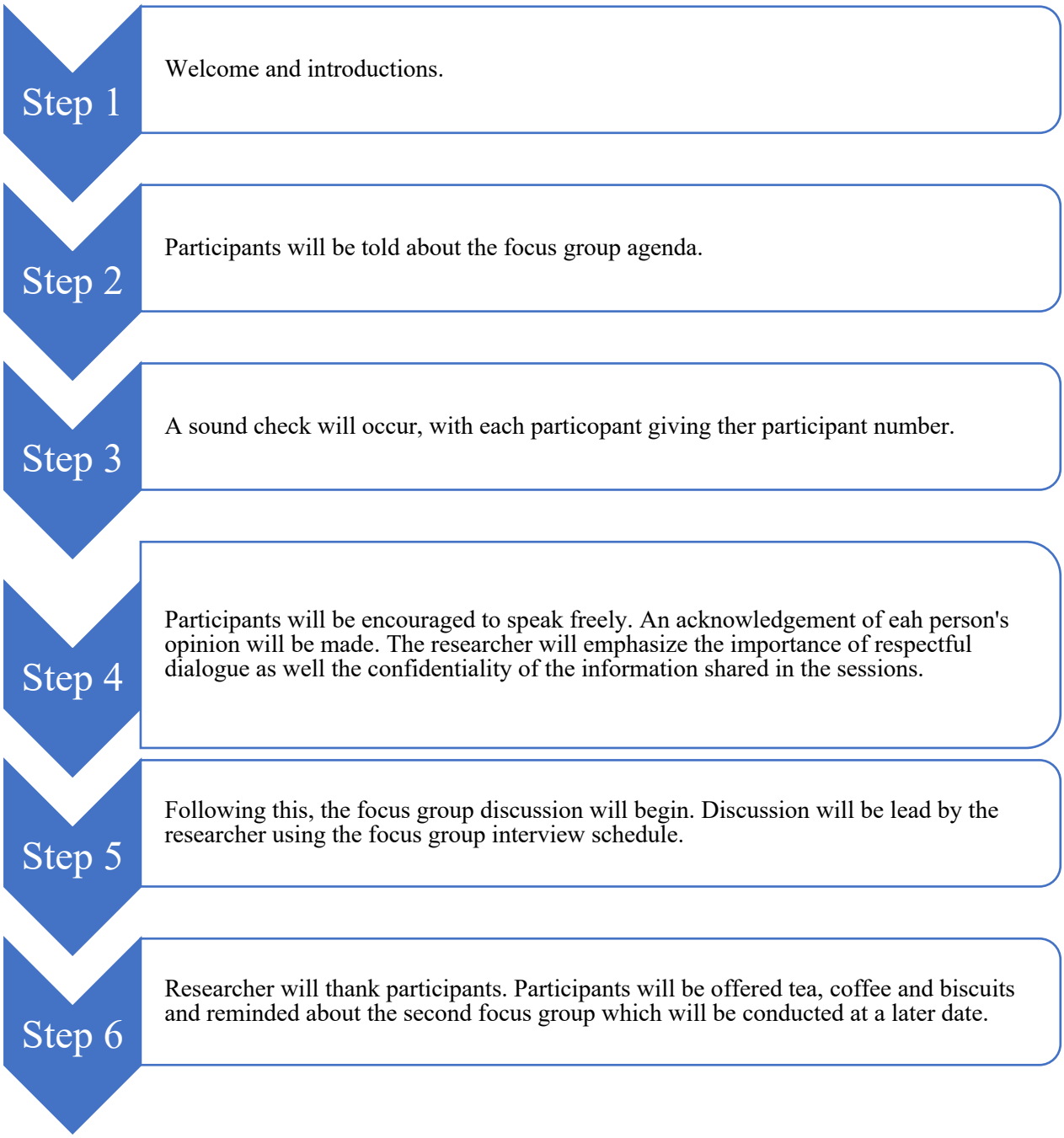
Declaration	**Yes**	**No**
I have read through the information sheet and understand it's content		
I understand that my consent is required		
I understand that participation is voluntary, and I can withdraw from the study without any consequences		
I understand that I will not be personally identified should this research study be published		
I consent to participating in this research study of my own free will		
I understand that I will not be paid for my participation		
I consent to having the focus group discussion recorded		

Participant signature: _____
Date: _____/_____/_____

Researcher signature: _____
Date: _____/_____/_____

APPENDIX J: Focus group interview schedules

1. Phase 1b caregiver group



A. Can you share an example of communicating with a child who uses AAC?
   - What experience have you had with children who use AAC devices that produce *a speech output*?
   - Tell me about the AAC devices/equipment that you currently use.

B. What kind of synthetic voices have you heard on AAC devices?
   - What have you noticed about the accents and languages on the devices?

- What have you noticed about the age of the voices on the devices?
- What have you noticed about the sex/gender of the voices on the devices?

C. What would you like the synthesised voices to sound like?
   - What is more important to you: The age, sex, accent, language, or intelligibility/ understandability of the voice?

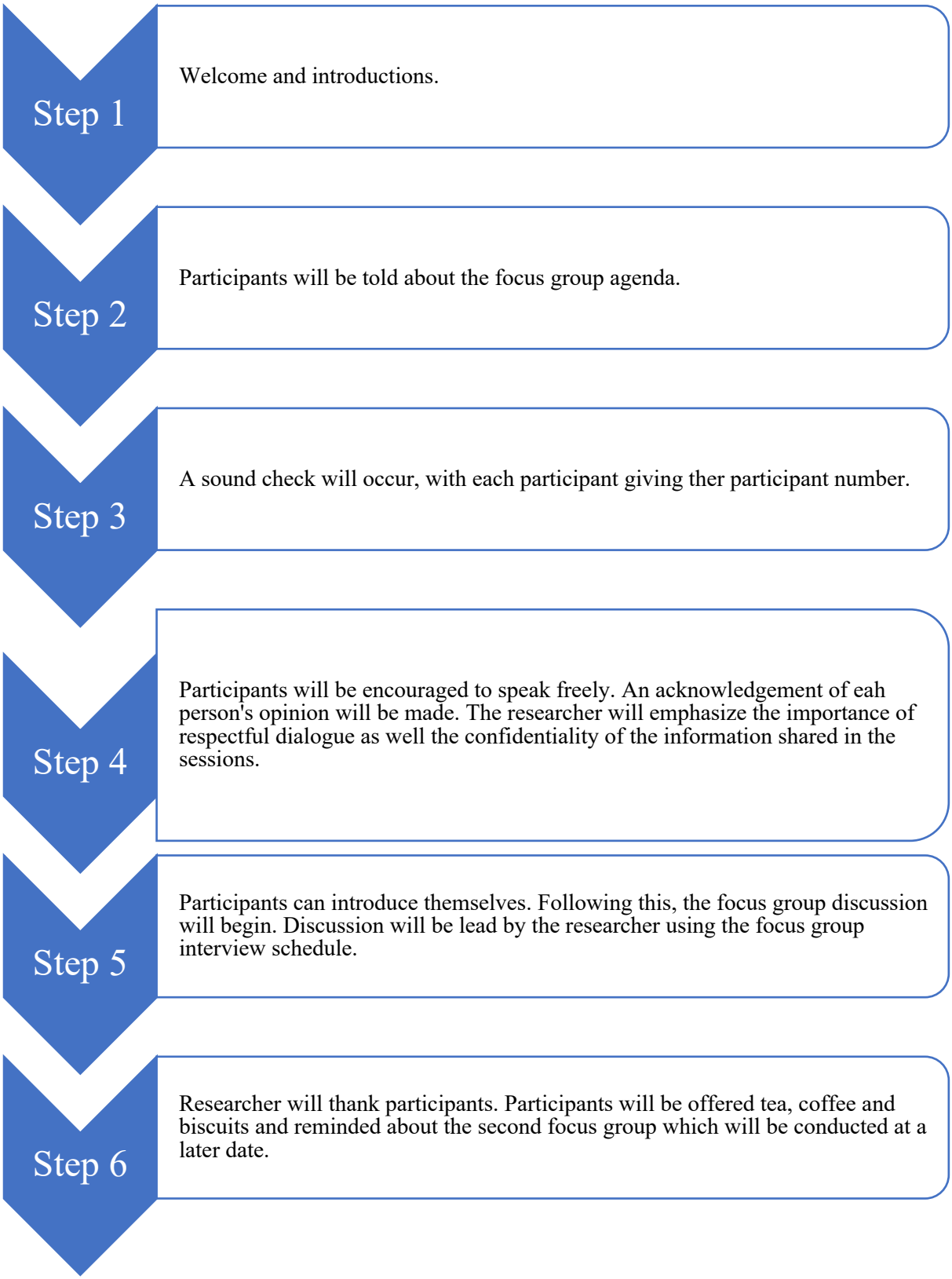
D. What makes a child stop using their AAC device?
   - What makes a parent/teacher stop using the child's AAC device to communicate with them?

E. How do you think other people view a child who uses an AAC device?
   - What kind of reaction would the community give if the children used a high-tech device? Why would they choose not to use a high-tech device?

F. In your opinion, what are the most important messages for the children to be able to communicate? (For example, basic needs)
   - Are there any words (personal, cultural, academic etc.) that you would want to add to the vocabulary?

G. What are the benefits of giving children a high-tech AAC device?
   - What could be some of the benefits in the home environment?
   - What could be some of the benefits in the school environment?
   - How would a *unique synthesised voice* make a difference for children?

2. Phase 1b professional's group (SLTs and teachers)



H. What experience have you had with children who use AAC devices that produce *a speech output*?

- Tell me about the AAC devices/equipment that you currently use.
- What kind of training have you received?
- Is AAC something you would focus on as a therapist? Do teachers make use of the AAC devices in the classroom?

I. Can you tell me about some of the practical challenges of implementing AAC in South Africa, particularly high tech AAC?

J. Do you have any ideas on how we may be able to overcome some of these challenges?

K. What kind of synthetic voices have you heard on AAC devices?
   - What have you noticed about the accents and languages on the devices?
   - What have you noticed about the age of the voices on the devices?
   - What have you noticed about the sex/gender of the voices on the devices?
   - In what way is the language on the device important to your students/clients/you?

L. What would you like the synthesised voices to sound like?
   - What is more important to you: The age, sex, accent, language, or intelligibility/ understandability of the voice?

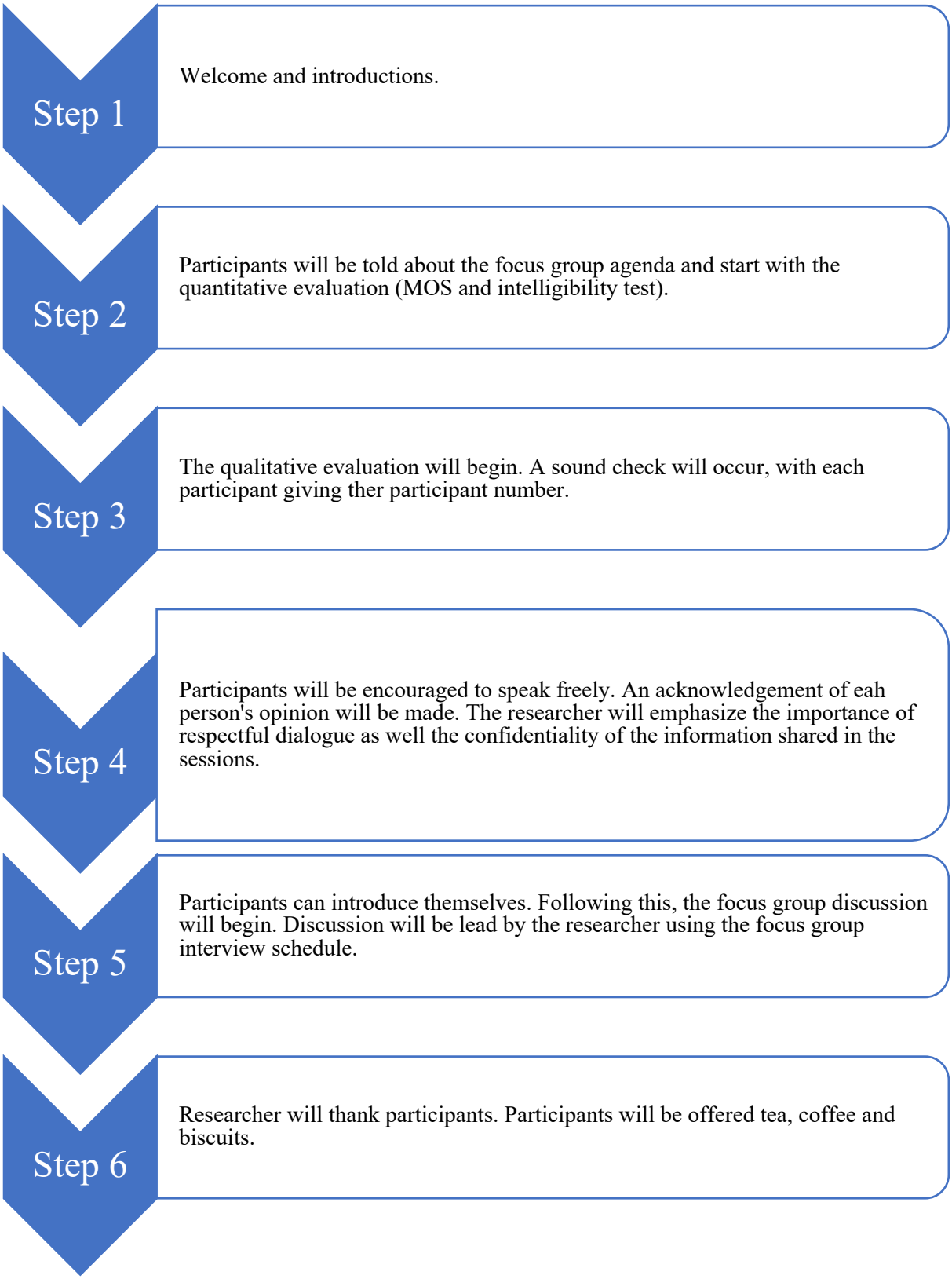
M. What makes a child stop using their AAC device?
   - What makes a teacher stop using the child's AAC device to communicate with them?

N. How do you think other people view a child who uses an AAC device?
   - What kind of reaction would the community give if the children used a high-tech device? Why would they choose not to use a high-tech device?

O. In your opinion, what are the most important messages for the children to be able to communicate? (For example, basic needs)
   - Are there any words (personal, cultural, academic etc.) that you would want to add to the vocabulary?

P. What are the benefits of giving children a high-tech AAC device?
   - What could be some of the benefits in the home environment?
   - What could be some of the benefits in the school environment?
   - How would a *unique synthesised voice* make a difference for children?

3. Phase 3 caregiver group



The focus group was divided into three sections. The first section involved participants listening to audio clips and giving their MOS. Participants listened to and evaluated eight unique synthetic speech

audio clips per language from one iPad with a loudspeaker. Following this, if participants were fully proficient in the language, they were asked to participate in intelligibility tests and transcribe six sentences per language, divided equally between synthetic child and adult speech. After a short break, the third section included the focus group discussion.

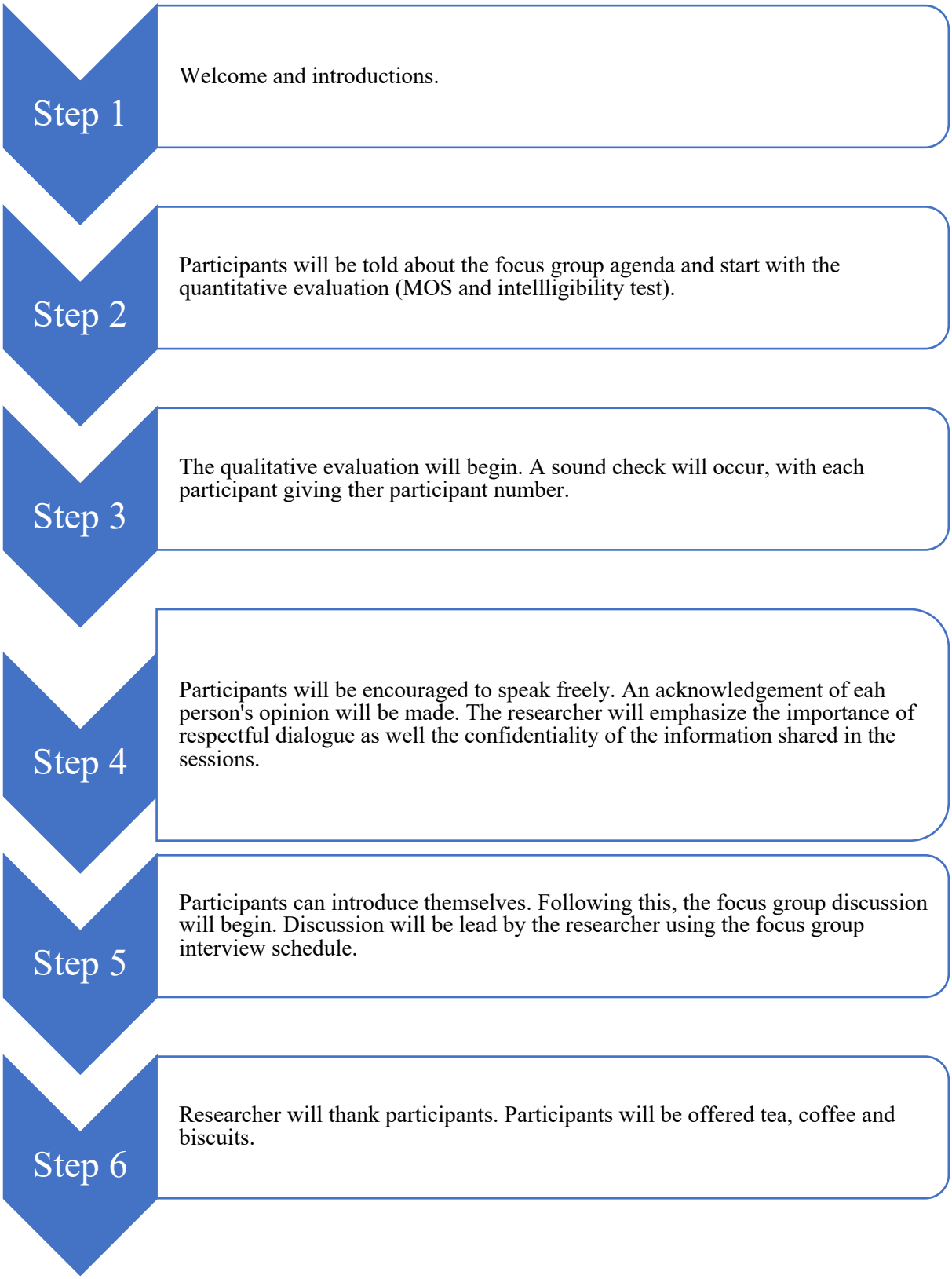
Focus group questions:

1. What do you think about the voices that you have just heard?
    - What do you notice about the accent, language, gender, age of the voices?
2. (Give example): If you had to choose between the old voices and the new child voices that we made, which would you prefer for the children, and why?
3. (Give example): If you had to choose between the child voices and the adult voices that we made, which would you prefer for the children, and why?
4. What, if anything, would you change about the child voices?

---

5. How would you communicate with the child if they had this iPad?
    - Effort, comfortable, willing to learn?
    - Support them using the iPad?
6. How do you think your family will view the child if they use the iPad with **these voices?**
7. How do you think the iPad could be used at school?
8. Why do you think your children might stop using the iPad?
    - Dislike, training, stigma, rejection/acceptance?

4. Phase 3 professional's group (SLTs and teachers)



The focus group was divided into three sections. The first section involved participants listening to audio clips and giving their MOS. Participants listened to and evaluated eight unique synthetic speech

audio clips per language from one iPad with a loudspeaker. Following this, if participants were fully proficient in the language, they were asked to participate in intelligibility tests and transcribe six sentences per language, divided equally between synthetic child and adult speech. After a short break, the third section included the focus group discussion.

Focus group questions:

1. What do you think about the voices that you have just heard?
    - What do you notice about the accent, language, gender, age of the voices?
2. (Give example): If you had to choose between the commercially available voices and the child voices that we made, which would you prefer for the children, and why?
3. (Give example): If you had to choose between the child voices and the adult voices that we made, which would you prefer for the children, and why?
4. What, if anything, would you change about the child voices?

---

5. How would you support a child using this iPad in the class?
    - Previous experience/ willing to learn, effort, comfortable, change teaching/ intervention plan?
6. How do you think the school community will view and interact with the child if they use the iPad with **these voices?**
    - Peers, support staff, teachers
7. Why do you think children might stop using the iPad to communicate?
    - Dislike, training, stigma, rejection/acceptance?

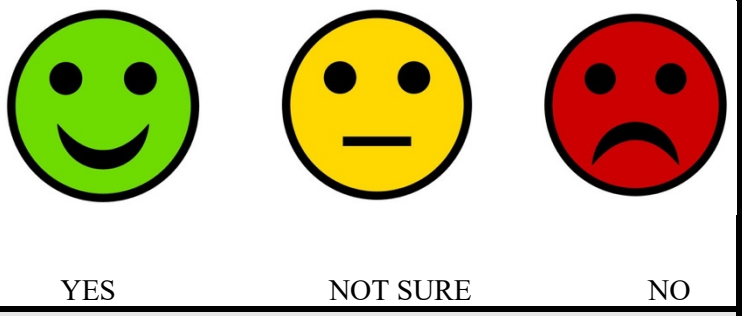
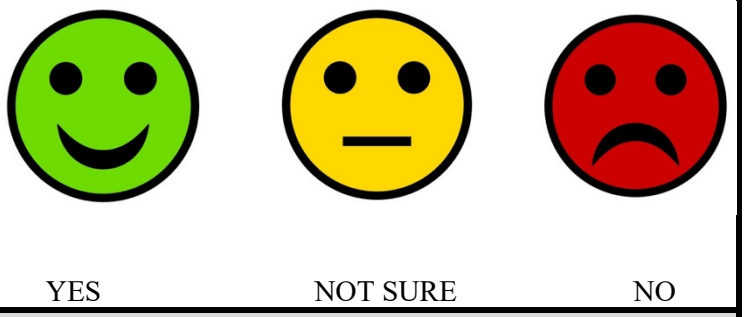
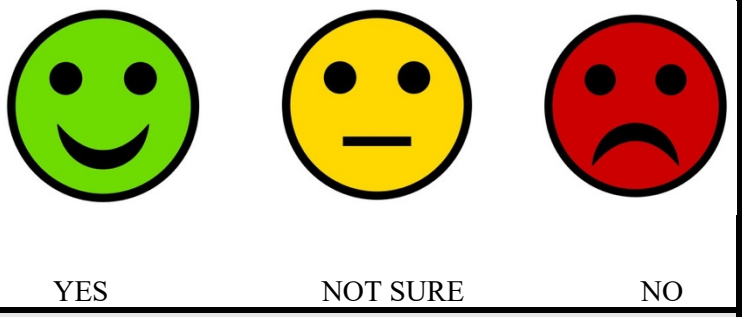
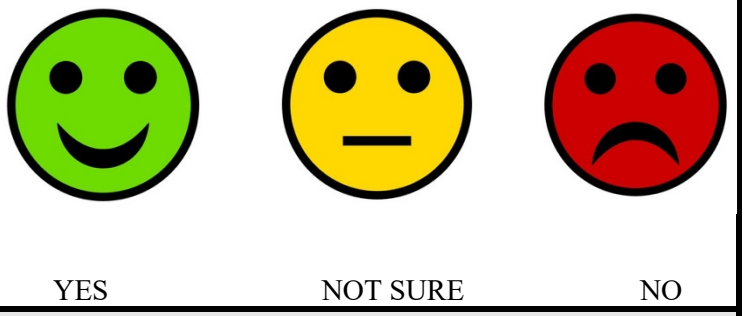
APPENDIX K: Speech synthesis evaluation sheets

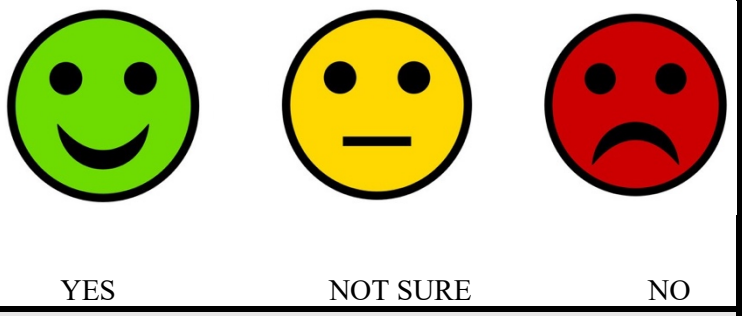
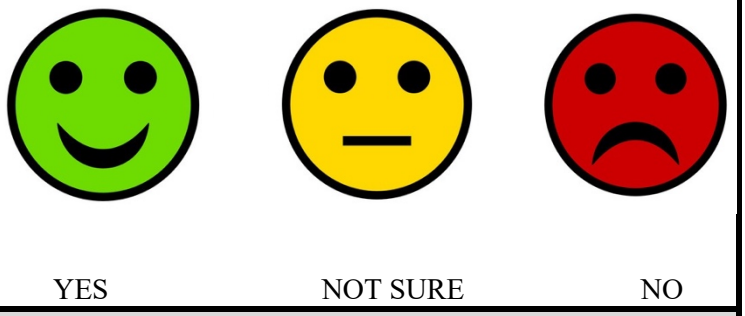
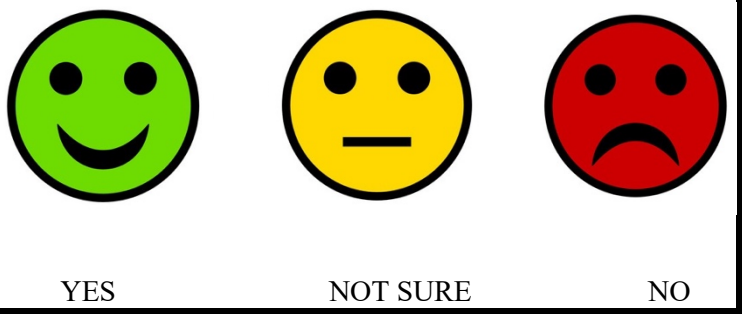
1. AAC-based evaluation sheet for children with CCN in Phase 3 (English example)

Each child listened to and provided feedback on seven synthetic speech audio clips in their home language, played from an iPad with a loudspeaker. This selection included three clips of synthetic child voices and, when asked to choose between adult and child voices, they were presented with a further two clips of each type.

Child name:

1	I will ask you some questions, is that okay?	 YES NOT SURE NO
2	Is your name ___?	 YES NOT SURE NO
3	Do you go to ____ school?	 YES NOT SURE NO

4	Do you like the child voice?	 YES NOT SURE NO
5	Do you think the voice sounds like you?	 YES NOT SURE NO
6	Does the language sound right?	 YES NOT SURE NO
7	Do you think you would use this voice to help you talk at school?	 YES NOT SURE NO

8	Do you think you would use this voice to help you talk at home?	 YES NOT SURE NO
9	Will people be able to understand you if you use this voice?	 YES NOT SURE NO
10	Which voice do you like the most? Which voice would you like to use?	(use stimuli below)
11	Do you want to tell me something else about the voice?	 YES NOT SURE NO

Stimuli for question 10:

2. MOS and intelligibility evaluation sheet for caregivers and professionals in Phase 3

Adult participants listened to and evaluated 14 unique synthetic speech audio clips per language from one iPad with a loudspeaker. This included two synthetic child audio clips and two synthetic adult audio clips to rate the overall impression, pleasantness, naturalness and understandability, and two audio clips for assessing similarity to real speakers (comparing to donor recordings). Following this, participants listened to six semantically predictable sentences per language (three synthetic child voices and three synthetic adult voices) for the intelligibility test.

Language: CHILD				
Overall impression	**Pleasantness**	**Naturalness**	**Understandability**	**Similarity**
"How do you rate the quality of the overall system?"	*"How would you describe the pleasantness of the voice?"*	*"How would you rate the naturalness of the voice?"*	*"How much listening effort was needed to understand what was said?"*	*"How similar is the synthetic voice to the real child voice?"*
1. Horrible	1. Very unpleasant	1. Very unnatural	1. Cannot understand	1. Very different
2. Poor	2. Unpleasant	2. Unnatural	2. Maximum effort needed	2. Slightly different
3. Tolerable	3. Satisfactory	3. Satisfactory	3. Fair	3. Average similarity
4. Good	4. Pleasant	4. Natural	4. Minimum effort needed	4. Almost identical
5. Excellent	5. Very pleasant	5. Very natural	5. No effort needed	5. Identical

Language:				**ADULT**
Overall impression	**Pleasantness**	**Naturalness**	**Understandability**	**Similarity**
"How do you rate the quality of the overall system?"	*"How would you describe the pleasantness of the voice?"*	*"How would you rate the naturalness of the voice?"*	*"How much listening effort was needed to understand what was said?"*	*"How similar is the synthetic voice to the real adult voice?"*
1. Horrible	1. Very unpleasant	1. Very unnatural	1. Cannot understand	1. Very different
2. Poor	2. Unpleasant	2. Unnatural	2. Maximum effort needed	2. Slightly different
3. Tolerable	3. Satisfactory	3. Satisfactory	3. Fair	3. Average similarity
4. Good	4. Pleasant	4. Natural	4. Minimum effort needed	4. Almost identical
5. Excellent	5. Very pleasant	5. Very natural	5. No effort needed	5. Identical

Intelligibility Test: Please write what you hear

1.	
2.	
3.	
4.	
5.	
6.	

Intelligibility test transcript

a) English child

	Coding	**Transcript**	**Length of file (in words)**	**Syllables**
1	SAE_child_SA2	The intelligent businessman had a very successful business on fourth street in Cape Town.	14	23
2	SAE_child_SA1	The small grey mouse ran up the lions neck and nearly went straight into his big mouth.	17	19
3	SAE_child_NAE5	They covered their mouths with their hands.	7	8

b) English adult

	Coding	Transcript	Length of file (in words)	Syllables
1	SAE_adult2	The sun warmed him and before he knew it, he was drifting off to sleep on the soft grass.	19	21
2	SAE_adult4	The children see lots of funny things on their way home.	11	13
3	SAE_adult6	They see a dangerous snake.	5	7

c) Afrikaans child

	Coding	Transcript	Length of file (in words)	Syllables
1	AFR_child_NAE5	Sy het in 'n paleis gewoon.	6	8
2	AFR_child_NAE1	Toe hy wakker word, is hy baie honger, en hy gaan soek iets om te eet.	16	19
3	AFR_child_SA5	Ek moes jou wakker gemaak het.	6	8

d) Afrikaans adult

	Coding	Transcript	Length of file (in words)	Syllables
1	AFR_adult1	Sy het nog nooit sulke wonderlike musiek gehoor nie.	9	15
2	AFR_adult2	Sy het nie geweet wat om te doen nie.	9	10
3	AFR_adult3	Naby die rivier is daar 'n groot slang, wat die hele winter geslaap het.	14	19

e) isiXhosa child

	Coding	Transcript	Length of file (in words)	Syllables
1	XHO_child_NAE2	Babebasa umlilo ngenkukuma ukugxotha ingqele.	5	18
2	XHO_child_SA3	Amantombazana ayethwele iinyanda entloko.	4	16
3	XHO_child_SA1	Yayiyeyona mvakwemini eshushu kakhulu ngoMgqibelo kuDisemba.	6	23

f) isiXhosa adult

	Coding	Transcript	Length of file (in words)	Syllables
1	XHO_adult3	Yayilusuku olumnandi, olungaxakekanga ngasemlanjeni.	4	21
2	XHO_adult4	Kuhlala iinyoka ezininzi kwihlathi elikwezi ntaba.	6	18
3	XHO_adult5	Ubusika obuqhaqhazelisa amazinyo babusele budlulile.	5	23

APPENDIX L: Human Research Ethics Committee Approval (765/2021)

UNIVERSITY OF CAPE TOWN
Faculty of Health Sciences
Human Research Ethics Committee

Room 45 E-52-E-Floor- Old Main Building
Groote Schuur Hospital
Observatory 7925
Telephone [021] 406 6492
Email: hrec-submissions@uct.ac.za
Website: www.health.uct.ac.za/fhs/research/humanethics/forms

14 March 2022

HREC REF: 765/2021

Dr M Harty
Division of CSD
Health & rehab Sciences-OMB
Email: michal.harty@uct.ac.za
Student: c.terblanche04@gmail.com

Dear Dr Harty

PROJECT TITLE: FEASIBILITY OF INDIVIDUALISED SYNTHETIC SPEECH FOR CHILDREN WITH COMPLEX COMMUNICATION NEEDS IN THREE SOUTH AFRICAN LANGUAGES (SOUTH AFRICAN ENGLISH, AFRIKAANS AND ISIXHOSA)
PHD CANDIDATE-MISS CAMRYN TERBLANCHE

Thank you for your response letter, addressing the issues raised by the Faculty of Health Sciences Human Research Ethics Committee (HREC).

It is a pleasure to inform you that the HREC has **formally approved** the above-mentioned study.

This approval is subject to strict adherence to the HREC recommendations regarding research involving human participants during COVID -19, our letter dated 02 February 2022 provides guidance found on our website:
http://www.health.uct.ac.za/fhs/research/humanethics/forms

Approval is granted for one year until the 30 March 2023.

Please submit a progress form, using the standardised Annual Report Form if the study continues beyond the approval period. Please submit a Standard Closure form if the study is completed within the approval period.
(Forms can be found on our website: www.health.uct.ac.za/fhs/research/humanethics/forms)

The HREC acknowledge that the student: -Miss Camryn Terblanche will also be involved in this study.

Please quote the HREC REF 765/2021 in all your correspondence.

Please note that the ongoing ethical conduct of the study remains the responsibility of the principal investigator.

Please note that for all studies approved by the HREC, the principal investigator **must** obtain appropriate institutional approval, where necessary, before the research may occur.

Yours sincerely

PROFESSOR M BLOCKMAN
CHAIRPERSON, FACULTY OF HEALTH SCIENCES HUMAN RESEARCH ETHICS COMMITTEE
Federal Wide Assurance Number: FWA00001637. Institutional Review Board (IRB) number: IRB00001938 NHREC-registration number: REC-210208-007
This serves to confirm that the University of Cape Town Human Research Ethics Committee complies to the Ethics Standards for Clinical Research with a new drug in patients, based on the Medical Research Council (MRC-SA), Food and Drug Administration (FDA-USA), International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use: Good Clinical Practice (ICH GCP), South African Good Clinical Practice Guidelines (DoH 2020), based on the Association of the British Pharmaceutical Industry Guidelines (ABPI), and Declaration of Helsinki (2013) guidelines. The Human Research Ethics Committee granting this approval is in compliance with the ICH Harmonised Tripartite Guidelines E6: Note for Guidance on Good Clinical Practice (CPMP/ICH/135/95) and FDA Code Federal Regulation Part 50, 56 and 312.

UNIVERSITY OF CAPE TOWN

FACULTY OF HEALTH SCIENCES
Human Research Ethics Committee

FHS016: Annual Progress Report / Renewal

HREC office use only (FWA00001637; IRB00001938)			
This serves as notification of annual approval, including any documentation described below.			
☑ Approved	Annual progress report	Approved until/next renewal date	30.08.2025
☐ Not approved	See attached comments		
Signature Chairperson of the HREC/ Designee		Date Signed	5/3/2024

Note: Please email this form and supporting documents (if applicable) in a combined pdf-file to hrec-enquiries@uct.ac.za.
Please clarify your plan for research-related activities during COVID-19 lockdown.
Please use the latest form found on our website:
http://www.health.uct.ac.za/fhs/research/humanethics/forms

HUMAN RESEARCH ETHICS COMMITTEE
- 5 MAR 2024
HEALTH SCIENCES FACULTY
UNIVERSITY OF CAPE TOWN

Comments to PI from the HREC

Principal Investigator to complete the following:

1. Protocol information

Date (when submitting this form)	4 March 2024		
HREC REF Number	765/2021	Current Ethics Approval was granted until	30 March 2024
Protocol title	Feasibility of individualised synthetic speech for children with complex communication needs in three South African languages (South African English, Afrikaans, and isiXhosa).		
Protocol number (if applicable)			
Are there any sub-studies linked to this study?		☐ Yes	✓ No
If yes, could you please provide the HREC Reference number for all sub-studies? **Note:** A separate FHS016 must be submitted for each sub-study.			
Principal Investigator	Michal Harty		
Department / Office Internal Mail Address	Michal.harty@uct.ac.za		

(Note: Please complete the Closure form (FHS010) if the study is completed within the approval period)

1.1 Does this protocol receive US Federal funding?	☐ Yes	✓ No
1.2 If the study receives US Federal Funding, does the annual report require full committee approval? **Note:** Any annual approvals for **Full Committee** review MUST be submitted on the monthly HREC submission dates. (Please send electronic copy for full committee review to hrec-submission@uct.ac.za)	☐ Yes	☐ No

If yes in 1.2 please complete section 1.3 below for invoicing purposes

1.3 Ethics Renewal Fee

Please **(tick ✓)**appropriate box for billing purposes:

Submission Type	*Description*	*New fee (Vat Incl.)*	*tick ✓*
Research funded solely from UCT departmental/ divisional/group budget	Annual evaluation of research progress report for re-certification	**R0,00**	✓
Non-sponsored student research for degree purposes at UCT/Other Universities & Colleges	Annual evaluation of research progress report for re-certification	**R0,00**	☐
Annual re-certification / Progress report (FHS016 Form)	Clinical Trial & International Grant Funded Research - Annual evaluation of research progress report for re-certification for Full Committee Approval	**R7000,00**	☐
Annual re-certification / Progress report (FHS016 Form)	Clinical Trial & International Grant Funded Research - Annual evaluation of research progress report for re-certification for Expedited review	**R3 710.00**	☐
Annual re-certification / Progress report (FHS016 Form)	National grant funded research - Annual evaluation of research progress report for re-certification for Full Committee Approval	**R6000.00**	☐
Annual re-certification / Progress report (FHS016 Form)	National Grant funded research for Annual evaluation of research progress report for re-certification for Expedited review	**R1 500,00**	☐

NB: Protocols funded by UCT (e.g. departmental funding / student research) and by certain grant funding organizations (e.g. MRC, NRF, CANSA,) are exempt from these charges.

Please provide details for Invoicing, either complete section 1 or 2 :

1. Invoice billing – Directly to Sponsor

Sponsor's name	NRF sponsored student
Billing Address of Sponsor:	NA
Vat Number:	NA
Contact person	NA
Telephone number	NA

Email Address	NA
2. **Internal Journal Billing:**	
Fund Number:	NA
Cost Centre Number:	NA
Account Holder Name:	NA
Division of Account Holder:	NA

2. List of documentation for approval

3. Protocol status (tick ✓)

✓	Open Enrolment
☐	Closed to enrolment (tick ✓)
☐	Research-related activities are ongoing
☐	Research-related activities are complete, long-term follow-up only
✓	Research-related activities are complete, data analysis only
☐	Main study is complete but sub-study research-related activities are ongoing
☐	Study is closed → Please submit a Study Closure Form (FHS010)

4. Enrolment

Number of participants enrolled to date	- 3 children with complex communication needs (phase 1b) - 6 adult family members (phase 1b) - 7 teachers and speech therapists (phase 1b) - 98 typically developing children (phase 2) - 123 adult listeners (phase 2)
Number of participants enrolled, since last HREC Progress report (continuing review)	237
Additional number of participants still required	0

5. Refusals

Total number of refusals (participants invited to join the study, but refused to take part)	0

6. Cumulative summary of participants

(Note: Please complete the Closure form (FHS010) if the study is completed within the approval period)

Total number of participants who provided consent	**241**
Number of participants determined to be ineligible (i.e. after screening)	3
Number of participants currently active on the study	16
Number of participants completed study (without events leading to withdrawal)	221
Number of participants withdrawn at participants' request (i.e. changed their mind)	0
Number of participants withdrawn by PI due to toxicity or adverse events	0
Number of participants withdrawn by PI for other reasons (e.g. pregnancy, poor compliance)	1
Number of participants lost to follow-up. Please comment below on reasons for loss of follow-up.	0
Number of participants no longer taking part for reasons not listed above. Please provide reasons below:	
N/A	

7. Progress of study

Please provide a brief summary of the research to date including the overall progress and the progress since the last annual report as well as any relevant comments/issues you would like to report to the HREC:
Data collection for the study is complete. We have completed phase 1a and published a journal article with the results. Data has been collected and analysed for phase 1b, and a manuscript has been written. We are currently waiting for the final outcome of the journal submission. Manuscripts have been written for phase 2 and 3, and they are undergoing review. Depending on the outcome of the reviews, we may need to conduct further analysis of the data. We are hoping to submit the full PhD study in 2024.

8. Protocol violations and exceptions (tick ✓ all that apply)

✓	No prior violations or exceptions have occurred since the original approval
☐	Prior violations or exceptions have been reported since the last review and have already been acknowledged or approved
☐	Unreported minor violations that have occurred since the last review, as well as significant deviations not yet reported, are attached for review

9. Amendments (tick ✓ all that apply)

☐	No Prior amendments have been made since the original approval

✓	Prior amendments have been reported since the last review and have already been approved
☐	New protocol changes/ amendments are requested as part of this continuing review (See note below)

Note: If new protocol changes are being requested in this review, please complete an amendment form (FHS006).
Specific changes in the amended protocol and consent/assent forms must be **bolded**, *italicised* or tracked and all changes must include a rationale.

10. Adverse events

10.1 Please provide below or attach a narrative summary of serious adverse events and/ or unanticipated problems since the last progress report. Please indicate changes made to the protocol and informed consent document(s) as a result (if not already reported to the HREC). Please comment on whether causality to any study procedure or intervention could be established.
There have been no adverse events or unanticipated problems.

10.2 Have participants received appropriate treatment/ follow-up/ referral when indicated (e.g. in the case of abnormal or incidental clinical findings, distress or anxiety)?		
☐ Yes	☐ No	✓ Not applicable
If yes, please describe:		

11. Summary of Monitoring and Audit Activities (tick ✓)

11.1 Was this study monitored or audited by an external agency (e.g. SAHPRA, FDA)?		
☐ Yes	☐ No	✓ Not applicable

11.2 Did a Data and Safety Monitoring Board publish a report?		
☐ Yes	☐ No	✓ Not applicable

11.3 If yes, please identify the agency and attach a summary of the findings.					
Agency Name		Report attached	☐ Yes	☐ No	✓ Not applicable
		DSMB report attached	☐ Yes	☐ No	✓ Not applicable

11.4 Has there been any agency, institutional or other inquiry into non-compliance in this study, or any finding of non-compliance concerning a member of the research team?	
☐ Yes	✓ No
If yes, please explain:	

(Note: Please complete the Closure form (FHS010) if the study is completed within the approval period)

12. Level of risk (tick ✓)

12.1 In light of your experience of this research, please indicate whether the level of risk to participants has:	
☐	Increased
☐	Decreased
✓	Shown no change
If there has been a change, please explain:	

12.2 Please provide a narrative summary of recent relevant literature that may have a bearing on the level of risk.
This is a minimal risk study and is largely qualitative. Participants are observed, audio recorded and requested to answer questions about their perceptions and experiences. Participants are not provided with therapy or treatment of any kind.

13. Insurance

Please confirm that valid no fault insurance is still in place? **(tick ✓)**			
☐ Yes		✓ No (not applicable)	
If yes, please complete the following:			
Insurer's name:			
Policy no.		*Coverage Period:	
For UCT sponsored studies please liaise the Insurance office via fhs.sponsorship@uct.ac.za regarding the required documentation and information required obtain a renewed UCT No-fault Insurance Certificate.			

14. Statement of conflict of interest

Has there been any change in the conflict of interest status of this protocol since the original approval? (tick ✓)	
☐ Yes	✓ No
If yes, please explain and if necessary, attach a revised conflict of interest statement (Section #7 in the New Protocol Application Form FHS013):	

(Note: Please complete the Closure form (FHS010) if the study is completed within the approval period)

15. Signature

My signature certifies that the above is complete and correct.			
Signature of PI		Date	04/03/2024

(Note: Please complete the Closure form (FHS010) if the study is completed within the approval period)

APPENDIX M: Data coding sheet used in Phase 1a scoping review (with example).

	Authors	Year	Aims/purpose	Design	Language	Study population and sample size	Method	Key findings
				1. Qualitative 2. Quantitative		**1. Sample size 2. Sex 3. Age 4. Typically developing children/children with disabilities**	**1. Novel synthesis system 2. Commercial synthesis system 3. Review**	
1	Begnum, M; Flatebø Hoelseth, S; Johnsen, Britt; Hansen, Frank	2012	a) Could this strategy be successful for producing a high quality child synthesis based on available Norwegian HMM voices? b) Aimed at exploring it's appropriateness and fit for a specific user group (comparing adult and child voice output)	1. Qualitative: Questionnaires, interviews (talking maps interviews) and observations	Norwegian	1. Three children ("primary" testers) with severe expressive communication impairments were selected (was four, one excluded) to participate in the user tests, along with their parents, schools and teachers ("secondary" testers). 2. Male. 3. 6-13 years. 4. Children with severe expressive communication impairments	1. Novel: Prototype an artificial Norwegian child voice based on speaker adaptive adjustment of adult based HMM synthesis.	a) The current prototype has challenges related to intonation and pronunciation, noise levels and naturalness and volume levels. b) High quality may be more important than identity fit in some context, but in other cases a voice with the characteristics of a child is valued as more important than "strange" sounds and occasionally flawed intonation. One way to interpret the findings is that parents respond positively towards a voice they deem more fitting for the child's identity. However, if focusing on the practicalities of using the synthesis in demanding surroundings, teachers and parents respond negatively to the prototype. c) The prototype was generally perceived to be about 7 years old and experienced as fitting the children's identity and age.