USING TACOTRON 2 TO DEVELOP SYNTHETIC CHILD SPEECH IN SOUTH AFRICAN ENGLISH

Camryn Terblanche[1], Benjamin V Tucker[2], Tyler Schnoor[3] and Michal Harty[1]
[1]*University of Cape Town, South Africa,* [2]*Northern Arizona University, United States,*
[3]*University of Alberta, Canada*

trbcam001@myuct.ac.za/ benjamin.tucker@nau.edu/ tschnoor@ualberta.ca/ michal.harty@uct.ac.za

ABSTRACT

There is often a substantial mismatch between the synthetic voices used on augmentative and alternative communication devices, and the child user's natural voice in relation to age, gender, dialect, and language. This study aims to determine if it is possible to develop realistic child speech synthesis, using Tacotron 2, for South African English. Two hours of child speech data were manually collected from one 11-year-old male child. Following this, two existing adult models were used to "warm start" the child speech synthesis. Despite the limited child speech data, we were able to successfully create a synthesised child voice of adequate quality. When a warm start training procedure is utilised, even when there is an age and dialect mismatch, the amount of training time decreases, and the synthesised voice matches the vocal quality of the child donor's voice.

Keywords: augmentative and alternative communication (AAC), children, speech synthesis, Tacotron, text-to-speech.

1. INTRODUCTION

Children with complex communication needs do not develop their speech, language, and/or communication skills in a typical pattern, most often as a result of: a) neurological disorders (e.g., intellectual disability), b) genetic disorders (e.g., Down syndrome), or c) structural abnormalities (e.g., cleft palate) [1]. When natural speech is restricted, augmentative and alternative communication (AAC) can assist individuals with complex communication needs establish functional communication skills. AAC includes techniques, strategies, and pictorial or written symbols. Using either low- or high-tech AAC options, it can be used to supplement or if needed, replace an individual's natural speech [2]. Low-tech AAC is made up of basic communication devices (which are often paper-based) whilst high-tech AAC involves speech-generating devices such as: mobile, computer, and/or tablet-based technologies [2]. Due to advancements in technology, high-tech AAC has improved considerably in recent years. There are many commercially available speech-generating devices and mobile applications for children. However, the speech output on these devices does not always match the age, gender, personality [3], dialect, and language of the child user [4]. Tönsing *et al.* [4] state that speech-generating devices in South Africa are usually provided in English dialects that are not necessarily reflective of South African English children's speech. As AAC research and the technology developments are frequently conducted in high-income, mostly English-speaking countries, it is not surprising that US-accented English, is most often incorporated in commonly used speech-generating devices [4], [5]. It is not unusual to walk into a South African special needs classroom and see all the children making use of the same adult US-English voice. These children may have a way to communicate, but with a voice that doesn't reflect their linguistic and cultural diversity. Although an individual's right to communicate is often discussed in AAC research, the rights of individuals using AAC to communicate in whichever dialect or language they choose, has not received equal attention [4]. Thus, this study aims to determine if it is possible to develop realistic child speech synthesis, using Tacotron 2, for South African English (SAE).

Although South Africa's official language policy allows schools to select any of the 11 official languages for teaching and learning, English in education, rather than one of the African languages, is often favoured by the community [6]. Unfortunately, this means that children using a speech-generating device at school are likely using their second or third language to communicate [4]. Despite the obvious language barrier, these children also have to use synthetic voices that do not necessarily match their age, gender, or dialect. Although this is an established problem, personalised synthetic speech software is not without cost [3], [7]. Until AAC applications meet the communication needs of the child user, AAC use will likely remain limited, resulting in fewer opportunities for participation and interaction. Despite this, producing realistic synthetic child speech from text is challenging and this is largely due to the scarcity of usable child speech corpora. Further challenges are

experienced when collecting child speech data [8]. For instance, in comparison to adult speech, children's read speech is typically less fluent, their speech often includes multiple articulatory errors and as the recordings are usually conducted in schools rather than sound-attenuated booths, background noise is common [8]. When speech data is however available, Tacotron 2 [9], an open-source speech synthesis system, can produce natural synthetic speech, with high similarity to human speakers [10]. Tacotron 2 is an end-to-end neural network-based text-to-speech system that can be trained on text-to-audio pairs, without phonetic annotation. The system is also user-friendly, which means that it can be used by individuals with limited text-to-speech experience. Additionally, as Tacotron 2 allows for rich conditioning of attributes, such as speaker and language, data adaptation is possible [9].

2. METHOD

One 11-year-old typically developing male child was recruited to record a total of two hours of read speech in SAE. Each recording session was thirty minutes long, with breaks every 10-15min. Using a Zoom H1 Handy recorder (44100 Hz), the recordings were collected in a repurposed classroom. The data were manually marked up into short utterance chunks of ≤13 seconds, using a Praat textgrid [11]. If audio files are ≤13 seconds, Tacotron 2 trains relatively quickly. Fluent speech was required, so all false starts, disfluencies and misarticulations were removed. After data cleaning, there was 113.7min of speech that remained. The sound files were extracted from the textgrids, changed to mono, and downsampled to 22050 Hz. The data were then randomly divided into training (90%) and validation files (10%).

The default Tacotron 2 [9] architecture was used to create the child speech synthesis. Tacotron 2 [9] is made up of a recurrent sequence-to-sequence feature prediction network that maps character embeddings to mel-scale spectrograms. A "warm start" training procedure was implemented, meaning that the child speech data was trained over a pre-existing model. Essentially, the learned voice from the pre-existing model is removed, while the linguistic characteristics of that voice remain intact. Therefore, as linguistic characteristics from the pre-existing model can be transferred to other speakers, the new model's time until convergence is reduced, which makes it particularly useful when data is limited. In the current study, two differing warm start models were compared. In *warm start A*, the published pre-trained Tacotron 2 model (tacotron2_statedict.pt) from NVIDIA, trained on the LJ dataset [12] was used. The LJ dataset consists of short recordings from one female adult North American English speaker. In total, the LJ dataset is approximately 24 hours long [12]. For *warm start B*, the child model was trained on the cleaned, resampled (22050 Hz) SAE Lwazi III text-to-speech dataset [13]. The Lwazi III dataset is made up of short and long audio clips from one female adult SAE speaker. Although the SAE dataset contains mostly SAE, there are also small amounts of additional language data. This provides phone coverage of other languages [13]. Specifically, it contained 6.5 hours of SAE, but also contained Afrikaans (2 min), isiZulu (10 min), isiXhosa (12 min), Sepedi (5 min) and Setswana (4 min) data.

Figure 1 illustrates the process used to generate the synthesis systems for the SAE child speech. As an adult SAE text-to-speech model for Tacotron 2 isn't available for download, this first had to be created. During training, the full Lwazi III dataset, including the additional language data, was used. It was theorised that the additional phone coverage may improve the pronunciation of non-English names, surnames, street and building names. Additionally, due to the frequency of loan words between languages, it was suspected that improved pronunciation may be achieved if additional language data were included. Due to the comparably restricted Lwazi III dataset, a warm start was implemented using the pre-existing Tacotron 2 model from NVIDIA. After manually cleaning the adult data and segmenting longer recordings, training took approximately 5 days. Following the creation of the SAE adult synthesis, the respective SAE adult model was used to warm start the SAE child speech synthesis. Using the generated mel-scale spectrograms, the published WaveGlow [14] model was then used as a vocoder to synthesise time-domain waveforms. WaveGlow is a flow-based generative speech synthesis program [14].

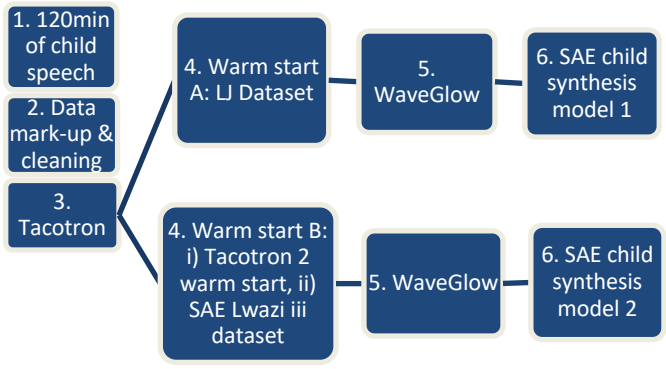

Figure 1: Process used for generating the synthesis systems for SAE child speech.

A further denoising step was included using a denoising code from the WaveGlow repository, which reduced background noise and eliminated

unwanted artefacts. In the end, naturalistic speech was created through a combination of Tacotron's prosody and WaveGlow's audio quality.

3. RESULTS

Encouragingly, despite the limited child speech data, we were able to successfully create a synthesised child voice of adequate quality in SAE. No matter which model was used to warm start the child synthesis, it took approximately 72 hours. It should be noted that the synthesised SAE adult voice was considerably less noisy than the two synthesised SAE child voices, but the additional language data used in the adult synthesis did not improve the pronunciation of non-English names and loan words to a great degree. Moreover, using a 5-point Likert Scale, where 0 was completely unnatural and 4 was completely natural, a mean opinion score method was used to gather 111 SAE listeners' (1998 responses) perceptions of the naturalness of the synthetic voices created. Results show that the adult speech ($\bar{x}$=3.38) was judged as more natural than the child speech ($\bar{x}$=2.20). Child synthesis model 2/warm start B ($\bar{x}$=2.33) was judged as more natural than model 1/warm start A ($\bar{x}$=2.06). The accent and vocal characteristics are comparable to the South African child donor, rather than the adult voice, no matter the adult model used. Figure 2 shows a spectrographic comparison between the donor child's speech and the speech output from the two SAE child models.

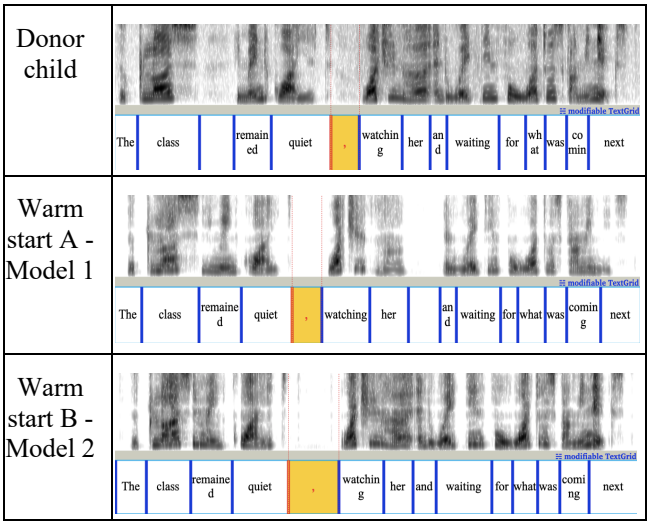

Figure 2: Spectrogram comparison between the SAE donor child and the synthesised child audio.

Alternatively, Figure 3 shows the mel-spectrograms and alignment plots for the two child models. The alignment plot is a useful tool to visualise 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 speech-like. The pacing is largely consistent, but there are occasional arbitrary pauses. The results also show that Tacotron 2 trains well with punctuation. E.g., a pause in speech due to a comma, is highlighted in Figure 2, and circled in Figure 3. However, warm start A produced some irregularities. E.g., the word *class* has a tin-like voice quality (arrow).

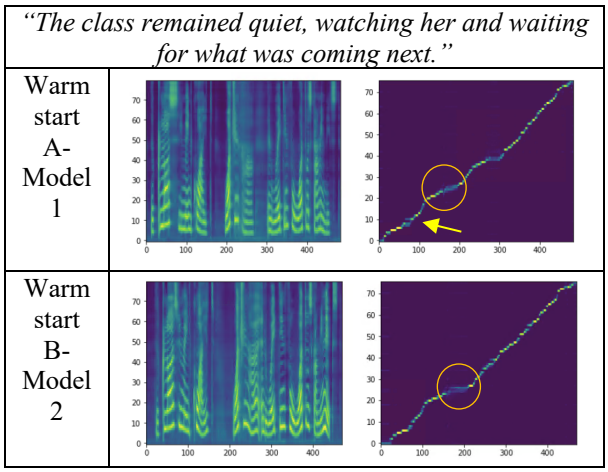

Figure 3: Comparison of the Tacotron 2 mel-spectrogram (a) and alignment (b) plots of the SAE synthesised child speech.

Model consistency was also considered. Figure 4 shows alignment plots between the two SAE child models (warm start A vs warm start B), when three sentences of differing lengths are synthesised.

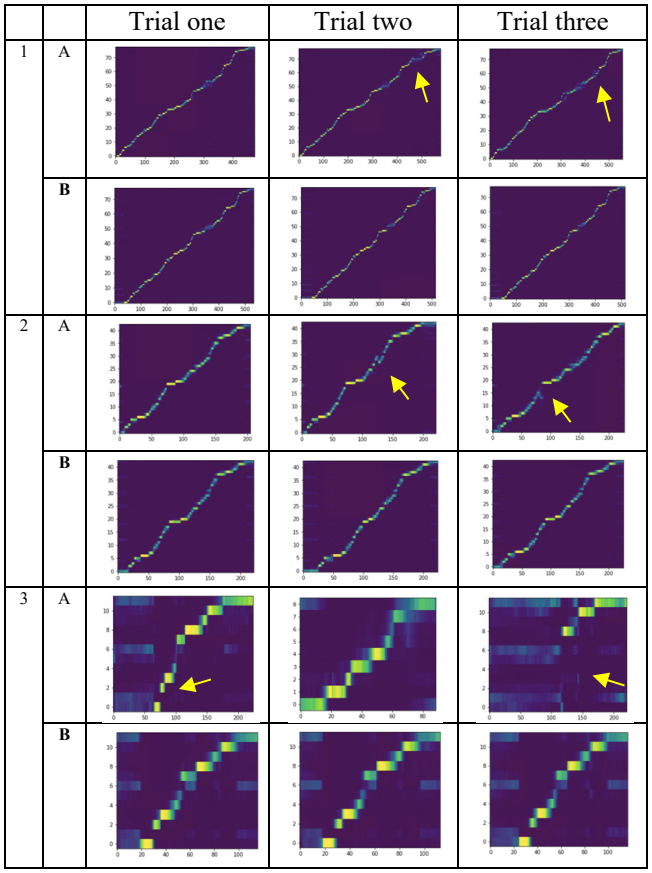

Figure 4: Alignment plots comparing the consistency of the two SAE synthesised child speech models.

It appears that the SAE child model is more consistent when warm start B is used, no matter the length of the sentence. In Figure 4, sentence 1 is long, *"Every day she woke up with a big smile on her face, because she loved her job."* Sentence 2 is medium length, *"They covered their mouths with their hands."* Sentence 3 is short, *"I like dogs."* The voices are mostly clear, but longer sentences produce better quality synthesis, as opposed to single-word or short sentences. Similarly to findings from Jain *et al.* [15], the first and last words in the phrases were more likely subject to distortions and artefacts, as compared to the middle of the phrase. The speaker's 'breathing' remained in the training data, which transferred to the synthetic models, making them appear more realistic.

4. DISCUSSION

Despite the occasional pronunciation and prosodic irregularity, adequate quality child speech synthesis was created with limited data. Both synthesised voices appear to sound like a relatively fluent South African child. Although the voices have minor distortions and some noise, the voices appear natural. As typically developing children are often hesitant when starting a phrase and may wander towards the end, it is unsurprising that these characteristics occasionally manifested in the synthetic model [15]. In fact, these characteristics likely contribute to greater naturalism as a typical child's speech patterns are less fluent, with variations in volume, pacing and emotional expressivity [15].

Firstly, it is effective to use a pre-trained North American English model to warm start a model in another English dialect. Secondly, using a warm start can greatly reduce the model's training time. This was observed when creating the SAE adult speech synthesis, and is in line with previous research [16], [17]. A warm start, using the Tacotron 2 model was utilised to create the SAE adult voice. The final SAE adult model (warm start B) was then utilised to warm start the SAE child synthesis. It was anticipated that including a warm start procedure twice, may result in the child model underperforming. Interestingly, the child model performed better with warm start B, suggesting that more than one warm start does not necessarily affect performance. Rather, the child model performance is directly proportional to the quality and the quantity of the adaptation data used. This is also true for the child speech data. In comparison, the adult model produced better synthesis quality, due to the data used. Although this study showed that only 113min of child data is sufficient, it is anticipated that if the donor child's speech were to be recorded in a sound-attenuated room, for an extended period, one would get improved child speech synthesis results.

Thirdly, the length of the text affects the speech output. Practically, this may affect the intelligibility of speech output on AAC devices. Drager [18] suggests that the context and length of the utterance play a role in the intelligibility of synthesised speech. Shorter sentences and single-word utterances are often less intelligible to the listener. Unfortunately, using a single-word AAC device is very common for new AAC users, and communication partners may find it difficult to understand the synthesised speech output, unless context is given. Fourthly, it is easier to find a high-quality adult corpus. This study has shown that by adapting adult data, using a warm start, one can successfully create a child voice that matches the vocal characteristics of the child donor. Thus, 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 model of high quality, can improve the quality of the synthesised child speech output. Although both warm start methods produced adequate quality child speech synthesis, if one were to use the voices for AAC purposes, it would be sensible to a choose a model with a consistent speech output. The fact that a warm start can be used when training data is limited, is beneficial for individuals who have complex communication needs. In the future, clinicians will be able to either collect the child's residual speech and incorporate it into the training model, or if speech is severely impaired, they will be able to use an age-matched typically developing child's voice. As Tacotron 2 [9] is open-source and produces naturalistic synthetic speech, the lack of extensive speech data should no longer limit the development of appropriate child voices. This technology can support marginalised AAC communities and develop much-needed resources in under-resourced languages.

Using adult adaptation data (warm start) to create a child voice is a viable method and could open the door for children with complex communication needs to have linguistically and culturally appropriate synthetic voices. As we successfully created a SAE child voice, future research should focus on creating child voices in other South African languages. It would also be interesting to determine if the residual speech of children with complex communication needs could be incorporated, so that the voice is a closer approximation to the child's natural voice [3]. Lastly, having children with complex communication needs select which voice they would prefer to use (i.e., an adult, a child, or a child voice that includes some of their own speech) is of interest to the authors.

5. ACKNOWLEDGMENTS

The financial assistance of the University of Cape Town and the National Research Foundation (NRF) of South Africa (DSINRF Reference Number: MND200619533947) is gratefully acknowledged. Additionally, this work was supported by Mitacs through the Mitacs Globalink Research Award Program. We also acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

6. REFERENCES

[1] M. Boesch and M. Da Fonte, *Effective Augmentative and Alternative Communication Practices: A Handbook for School-Based Practitioners*. Routledge, 2019.

[2] Y. Elsahar, S. Hu, K. Bouazza-Marouf, D. Kerr, and A. Mansor, 'Augmentative and Alternative Communication (AAC) advances: A review of configurations for individuals with a speech disability', *Sensors*, vol. 19, 2019.

[3] T. Mills, T. Bunnell, and R. Patel, 'Towards personalized speech synthesis for augmentative and alternative communication', *Augmentative and Alternative Communication*, vol. 30, no. 3, pp. 226–236, 2014.

[4] K. Tönsing, K. van Niekerk, G. Schlünz, and I. Wilken, 'Multilingualism and augmentative and alternative communication in South Africa – Exploring the views of persons with complex communication needs', *African Journal of Disability*, vol. 8, no. 0, 2019.

[5] C. Terblanche, M. Harty, M. Pascoe, and B. V. Tucker, 'A situational analysis of current speech-synthesis systems for child voices: A scoping review of qualitative and quantitative evidence', *Applied Sciences*, vol. 12, no. 5623, pp. 1–17, 2022.

[6] South African Government, 'Language in Education Policy Document (LiEP)'. Pretoria: Government Printers, 1997. Accessed: Oct. 21, 2022. [Online]. Available: https://www.education.gov.za/Portals/0/Documents/Policies/GET/LanguageEducationPolicy1997.pdf

[7] C. Jreige, R. Patel, and H. T. Bunnell, 'VocaliD: Personalizing Text-to-Speech Synthesis for Individuals with Severe Speech Impairment', *Assets '09*, pp. 259–260, 2009.

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

[9] Y. Wang *et al.*, 'Tacotron: Towards end-to-end speech synthesis', *arXiv*, 2017, [Online]. Available: https://doi.org/10.48550/arXiv.1703.10135

[10] X. Wang *et al.*, 'ASVspoof 2019: A large-scale public database of synthetic, converted and replayed speech', *arXiv*, 2020, doi: https://doi.org/10.48550/arXiv.1911.01601.

[11] P. Boersma and D. Weenink, 'Praat: Doing phonetics by computer [Computer program]'. 2022. [Online]. Available: http://www.praat.org/

[12] K. Ito and L. Johnson, 'The LJ speech dataset', 2017, [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[13] A. Louw and G. Schlünz, 'Lwazi III English TTS Corpus', vol. 1. Meraka Institute, CSIR, 2016. [Online]. Available: https://hdl.handle.net/20.500.12185/267

[14] R. Prenger, R. Valle, and B. Catanzaro, 'Waveglow: A flow-based generative network for speech synthesis', *arXiv*, 2018, [Online]. Available: https://doi.org/10.48550/arXiv.1811.00002

[15] R. Jain, M. Yiwere, Bigioi, P. Corcoran, and H. Cucu, 'A text-to-speech pipeline, evaluation methodology, and initial fine-tuning results for child speech synthesis', *arXiv*, 2022, doi: https://doi.org/10.48550/arXiv.2203.11562.

[16] E. Cooper *et al.*, 'Zero-shot multi-speaker text-to-speech with state of the art neural speaker embeddings', presented at the IEEE (4-8 May 2020), 2020.

[17] V. Barnekow, D. Binder, N. Kromrey, P. Munaretto, A. Schaad, and F. Schmieder, 'Creation and detection of German voice deepfakes', presented at the International Symposium on Foundations and Practice of Security, 2021, pp. 355–364.

[18] K. Drager, J. Reichle, and C. Pinkoski, 'Synthesized Speech Output and Children: A Scoping Review', *American Journal of Speech-Language Pathology*, vol. 19, pp. 259–273, 2010.