It’s A World Of Robots: Speech Synthesis And The Issue Of Everyone Having The Same Voice

UNIVERSITY OF ALBERTA
Department of Linguistics

Jasmine Wegewitz, Tyler Schnoor, Benjamin V. Tucker

Alberta Phonetics Laboratory

Why?

- The way we produce language is highly variable, multiple factors contribute to this (e.g. dialect, age, language experience).
- Despite language diversity those who rely on alternative communication are often limited to an adult voice and certain dialects. (Mills et al., 2014)
- Although a challenge; it is now possible to synthesize an intelligible child voice due to technology improvements. (Terblanche et al., 2022)

Research Questions:

- Is it possible with the resources available to successfully train a text to speech model to synthesize child speech?
- How much speech do we need?

How?

- One female nine-year-old donor recorded approximately ninety minutes of speech during three audio sessions (Figure 1).
- Data was marked up into utterance chunks (Figure 2).
- From there the marked up data was then used to train Tacotron2 (a deep neural network).

Procedure:

1. We used a sound attenuated booth to make the recordings and took precautions to reduce distractions and background noise.
2. Throughout the recordings we maintained the controlled variables: microphone placement, explanation, and breaks every ten-fifteen minutes or so in order to keep the voice fresh and the speaker happy.

Modeling:

- Trained two separate speech synthesis models using the default Tacotron architecture.
  - Model 1: 35 minutes of child speech
  - Model 2: 64 minutes of child speech
- We used a “warm start” training procedure which consisted of training child data over a pre-existing model in order to develop child synthesis.
  - Pre existing model: 20+ hours of adult speech
- We used a published WaveGlow model to generate a waveform from Tacotron’s mel spectrogram outputs.

Figure 1: The Recordings

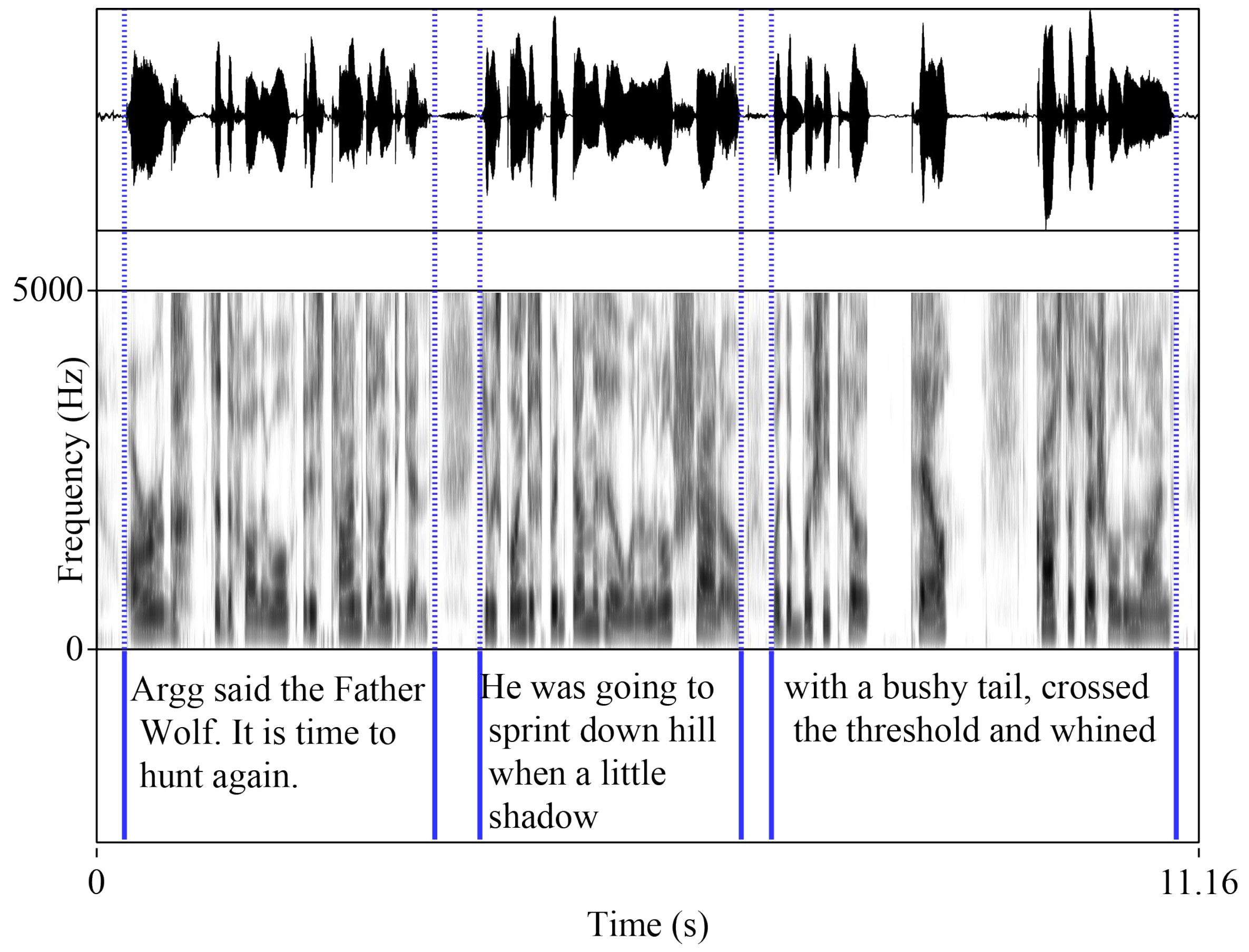

Figure 2: Spectrogram Mark up

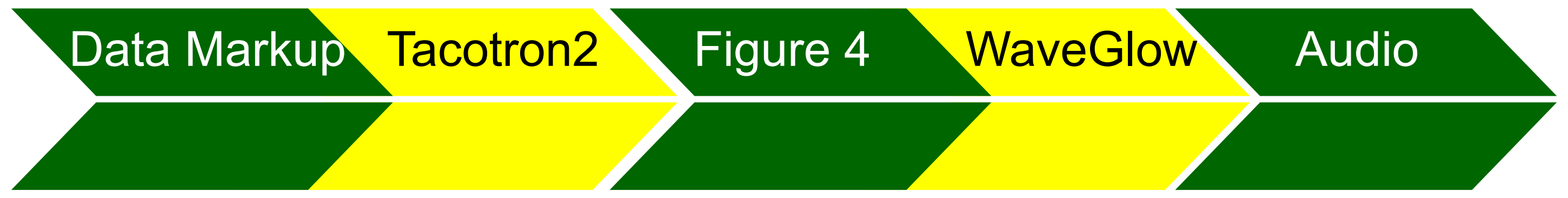

Figure 3a: Our process for synthesizing speech

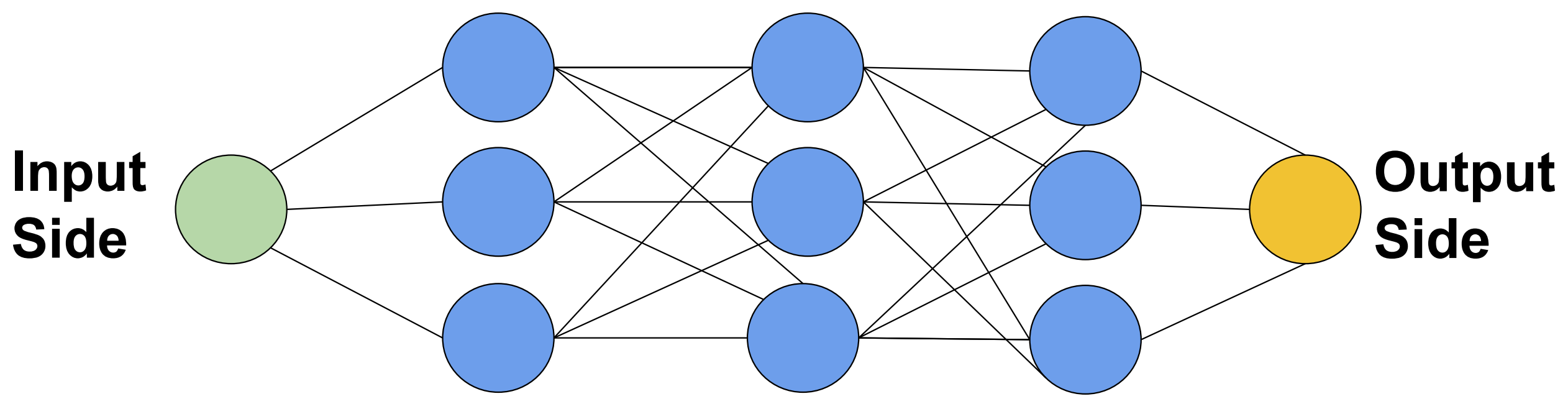

FIgure 3b: Neural Network

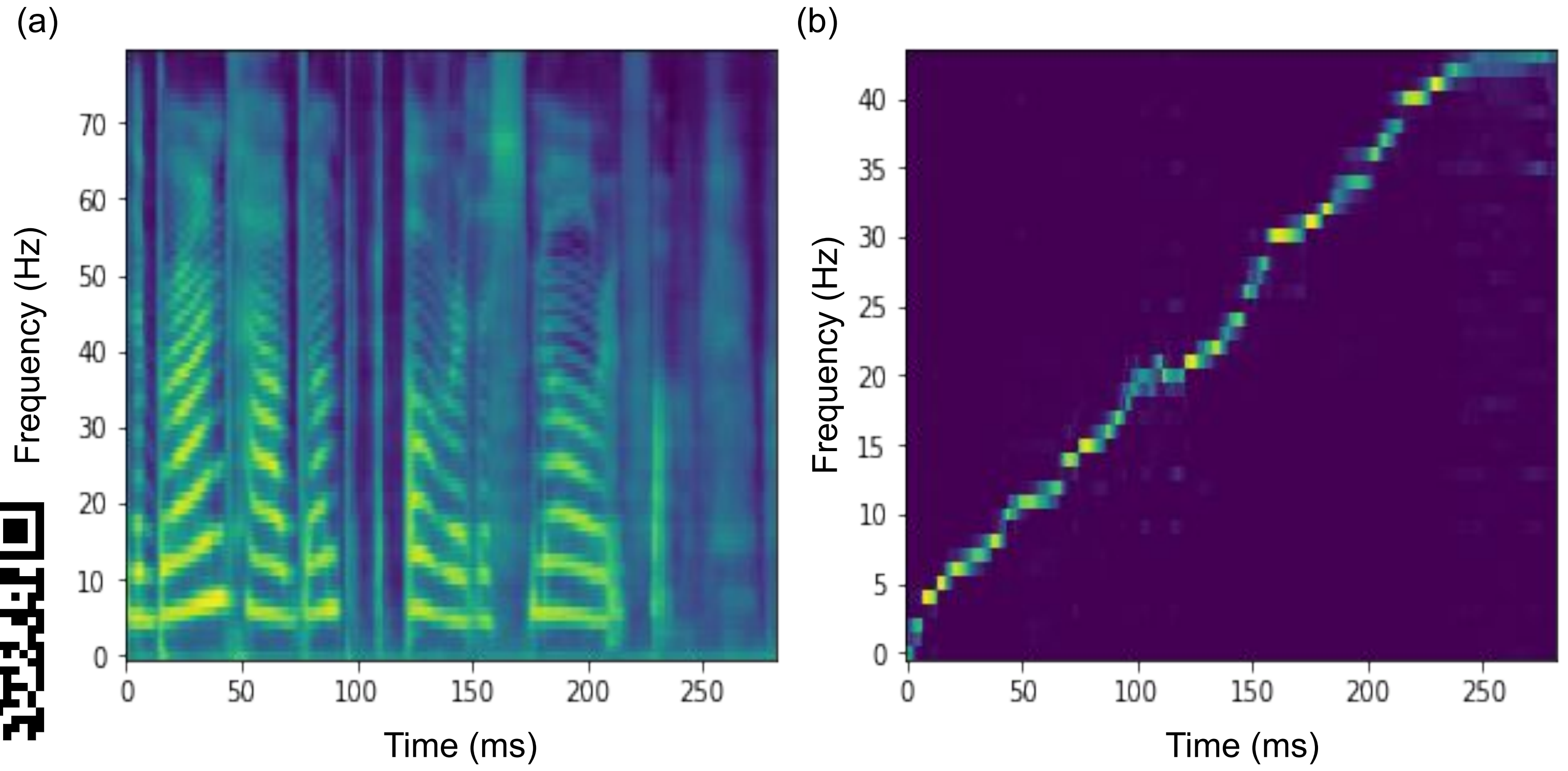

Figure 4 Tacotron 2 mel-spectrogram (a) and alignment (b) plots of synthesized speech

Results

- Analysis of the Deep Neural Networks proved to be very effective. As the data from the first recording was process the trend decreased from a 37% error to 12% error by the end of the first trial.
- 35 minutes of training material created a decently intelligible speech synthesis system
- The model with more data (64 minutes) had improved intelligibility in the synthesis.
- Figures 4 (b) displays a Tacotron 2 output from the first recording. The alignment plot presented indicates the quality of the model.

Qualitative Results:

- The first recording proved intelligible with low error with only a small sample of speech.
- The more data the DNN analyses and processes the better the model gets.

Conclusions/Impacts

- This method shows promise for child speech synthesis projects for future studies..

The future of speech synthesis

- The application of child speech technology can impact children who rely on alternative communication.
- The Tacotron 2 base model can be fine tuned on small speech samples.
- This method is applicable for generating a range of voices to better portray the speaker’s identity.

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. *Appl. Sci.* 12. 5623.

Shen, J. et al., (2018). Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions,*IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*,, pp. 4779-4783, doi: 10.1109/ICASSP.2018.8461368.

Wang, Y. et,. al. (2017). Tacotron: Towards End-To-End Speech Synthesis

Mills, T.; Bunnell, H.T.; Patel, R. (2014) Towards personalized speech synthesis for augmentative and alternative communication. *Augment. Altern. Commun. 30*, 226–236.

Acknowledgements: My Sarcastic Labmates Tori, Laura, Niki, Tyler, and Dr. Tucker who taught me that being an adult only means that you are old enough to buy the Dinosaur Chicken nuggets and the smiley fries from the store, and only Sith speak in absolutes. I would also like to thank my family you know who you are who have helped me every step of the way; I am exceedingly grateful. And to my mentors who have challenged me, and inspired me. I thank you all.