Universal Approach to Multilingual Multispeaker Child Speech Synthesis

Shaimaa Alwaisi, Mohammed Salah Al-Radhi, Géza Németh

Shaima.alwaisi@bme.hu.edu

1. Research Question

Why child speech important ?

- The challenges posed by limited child datasets .
- adds adult datasets during the training phase .
- Cross-Linguistic Variability.
- captures attention and keeps listeners involved.

2. Problem Formulation

- Individuality and Linguistic Diversity: Just as with adults, children have individual speech patterns influenced by factors like native language, accent, and anatomical differences:
- Ethical and Responsible Use: Ethical considerations surrounding the use of child speech synthesis should be part of the problem formulation, ensuring that the technology is used in ways that respect privacy, consent, and the well-being of children.

3. Goals

- Design child speech synthesis systems that can capture the unique speech characteristics of individual children.

5. Methods

- Propose approach based on modified AutoVocoder for synthesizing speech for multi-speakers, boys and girls.
- Explores the utilization of a universal neural vocoder based on BIGVGAN and AutoVocoder for synthsize child speech in both English and Hungrain languages.

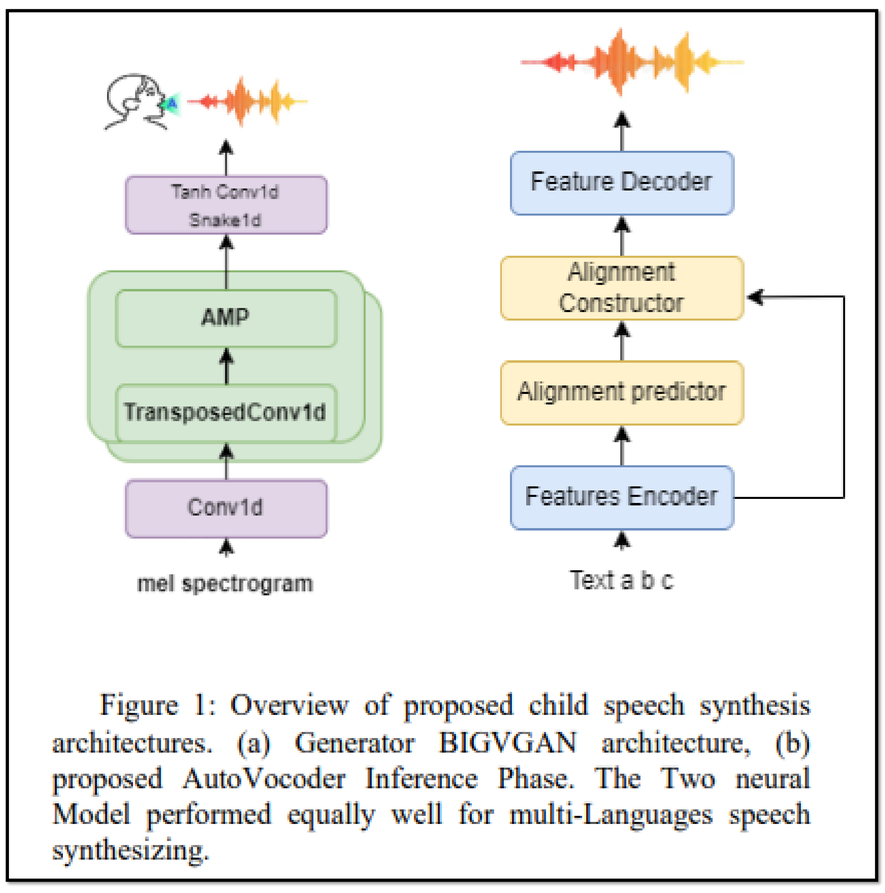

Figure 1: Overview of proposed child speech synthesis architectures. (a) Generator BIGVGAN architecture, (b) proposed AutoVocoder Inference Phase. The Two neural Model performed equally well for multi-Languages speech synthesizing.

6. Experimental conditions

Datasets

- My Science Tutor (MyST) consists of conversational child speech that is collected from 1371 students in grades ranging from third to fifth grade.
- We convert the speech sampling rate to 16KHz.
- For training the AutoVocoder, the English LJ speech1.1 dataset was used, which consists of recordings from a single female speaker. We used the female dataset since female voices share some acoustic characteristics with child voices.

Model

- The AutoVocoder was trained at a batch size of 16 and learning rate of 0.0002 and took three days of training to reach 210k Epochs, training process was executed on a system running Ubuntu 16.04.7 LTS, with NVIDIA-SMI and CUDA version 11.4 for efficient utilization of GPU resources.
- the AutoVocoder on the adult LJ speech 1.1 dataset and then fine tuned to low resources child datasets.

7. Results

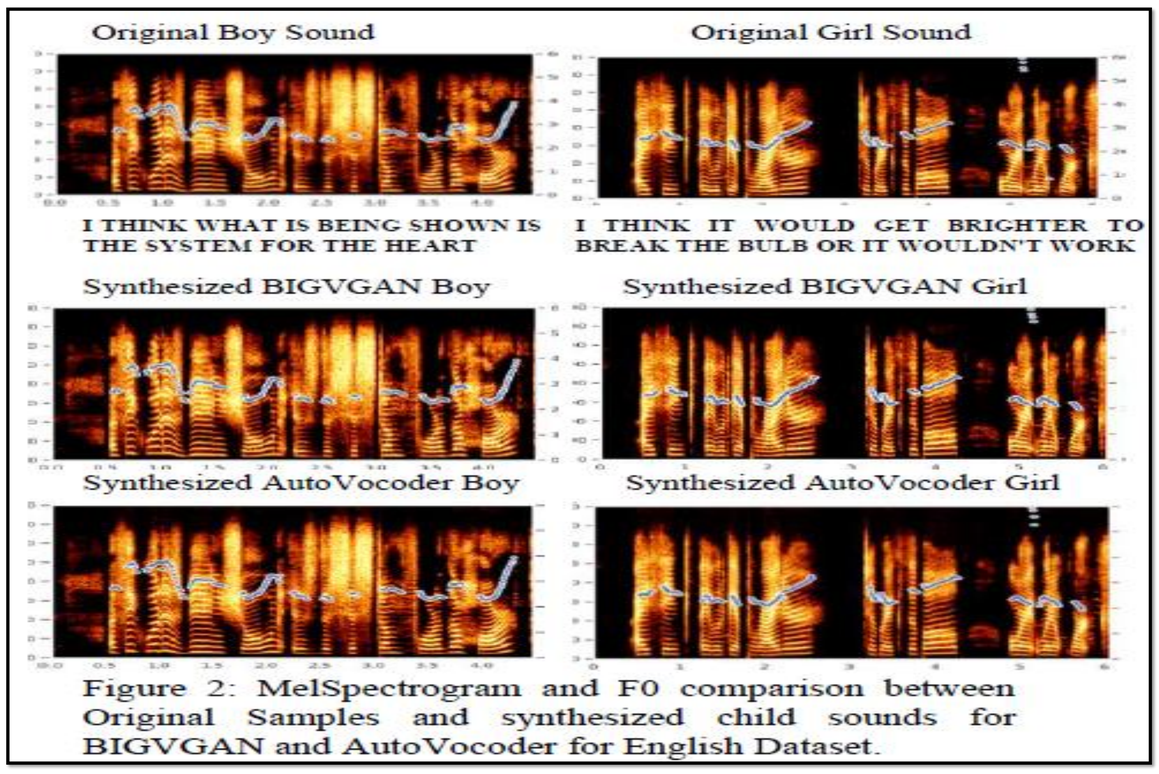

Figure 2: MelSpectrogram and F0 comparison between Original Samples and synthesized child sounds for BIGVGAN and AutoVocoder for English Dataset.

- Figure 2 demonstrates example of synthesized speech from the reference speaker.
- As observed that our model synthesizes high quality mel-spectogram which is comparable to ground-truth mel-spectrogram.
- Based on the losses, model is performing well on the training and validation data.

Table 1: *MOSNet Scores for English and Hungarian languages*

Samples	Boy		Girl	
	English	Hungarian	English	Hungarian
Original	3.292	2.921	2.899	2.675
BIGVGAN	3.228	2.847	2.925	2.661
AutoVocoder	3.652	2.923	3.041	2.738

8. Conclusion and Future work

- Propose speech synthesis models for multi child speech in English and Hungrian languages. The child speech synthesis models based on universal neural vocoder BIGVGAN and AutoVocoder.
- Our model's ability to generate high-quality spectrograms has been verified through experiments using reference speakers.
- We will evaluate subjective naturalness of synthesized speech
- We will extend the TTS model to more languages and Enable multi-lingual speech style transfer
- This research marks significant progress in child speech synthesis models for with potential to use the proposed AutoVocoder and Fastpitch to achieve TTS for children as further advancements in the future.

References

1. Webber, Jacob J., Cassia Valentini-Botinhao, Evelyn Williams, Gustav Eje Henter, and Simon King. "Autovocoder: Fast waveform generation from a learned speech representation using differentiable digital signal processing." International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5. IEEE, 2023.
2. Lee, Sang-gil, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. "Bigvgan: A universal neural vocoder with large-scale training." arXiv preprint arXiv:2206.04658 (2022).
3. Ward, Wayne, Ron Cole, and Sameer Pradhan. "My science tutor and the myst corpus." Boulder Learning Inc (2019).

TMiT SmartLab Intelligent Interactions

Budapest University of Technology and Economics
Department of Telecommunications and Media Informatics
Budapest, Hungary

MŰEGYETEM 1782