Multi-Speaker Child Speech Synthesis in Low-Resource Hungarian Language

Shaimaa Alwaisi, Mohammed Salah alradhi, Géza Németh
BME TMIT SmartLabs

Workshop on Intelligent Infocommunication Networks, Systems and Services

Speaker: Shaima Alwaisi

MŰEGYETEM 1782

http://smartlab.tmit.bme.hu

SmartLab
Intelligent Interactions

TMiT

CONTENTS

01 INTRODUCTION

02 METHODOLOGY

03 EXPERIMENTAL RESULTS

04 CONCLUSION

01

INTRODUCTION

Introduction and Related works

Introduction

Rapid advancements in TTS driven by DNN models have made TTS systems popular in various domains. However, designing TTS models for children is challenging due to distinct characteristics of child speech.

Child speech synthsis

Challenges in Child Speech TTS

Speech Characteristics

Mispronunciations, disfluencies, and linguistic differences set child speech apart from adult speech.

Acoustic Properties

Diferences in durations, pitch, and formant frequencies in children's speech

Data Challenge

the lack of sufficient child speech datasets.

Literature efforts

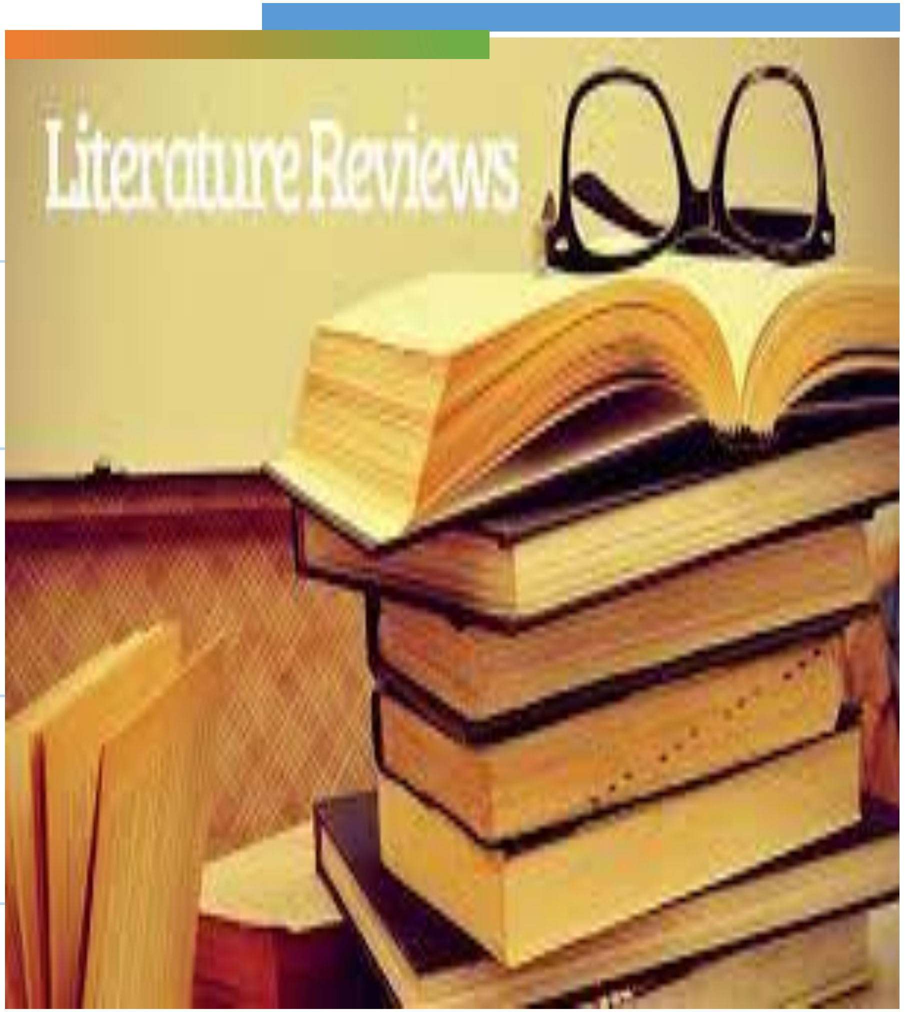
01 Early methods used concatenative, parametric, and HMM-based approaches

02 HMM-based systems struggled with naturalness.

03 Jain et al. used deep neural vocoders for child speech synthesis.

04 Commercial solutions like Acapela Group for child speech synthesis.

05 Hybrid DNN-HMM for Italian child speech recognition.



02

METHODOLOGY

Data Preparation, AutoChild speech synthesis Vocoder Details

Vocoders used in our study

01

AutoVocoder

We compared AutoVocoder with the BigVGAN in terms of synthesized sound quality

02

BigVGAN

BigVGAN with architecture and hyperparameters from literature. BigVGAN, we used a checkpoint of 500k iterations that are trained on the same English LJ speech1.1 dataset

METHODOLOGY

Datasets

Our experimentation is conducted on the GABI Hungarian child speech dataset GABI . we focused on records from children aged 7 to 11 years engaged in conversations. The speech samples were digitally recorded at 44.1 kHz, 16-bit

AutoChild speech synthesis

AutoVocoder adapted for child speech synthesis, enabling real-time applications and integration with various systems. Trained with specific settings and multi-language capabilities.

Data Preparation

Conversational Multi speaker child dataset

- we extracted a subset comprising wave files for ages 7 to 11 years old. Additional filtering was applied to select wave files with low background noise levels.
- we truncated long wave files to a maximum duration of 10 seconds
- GABI dataset presented challenges like noise and varying recording lengths; we filtered a subset for clarity.

Hungarian child speech dataset GABI

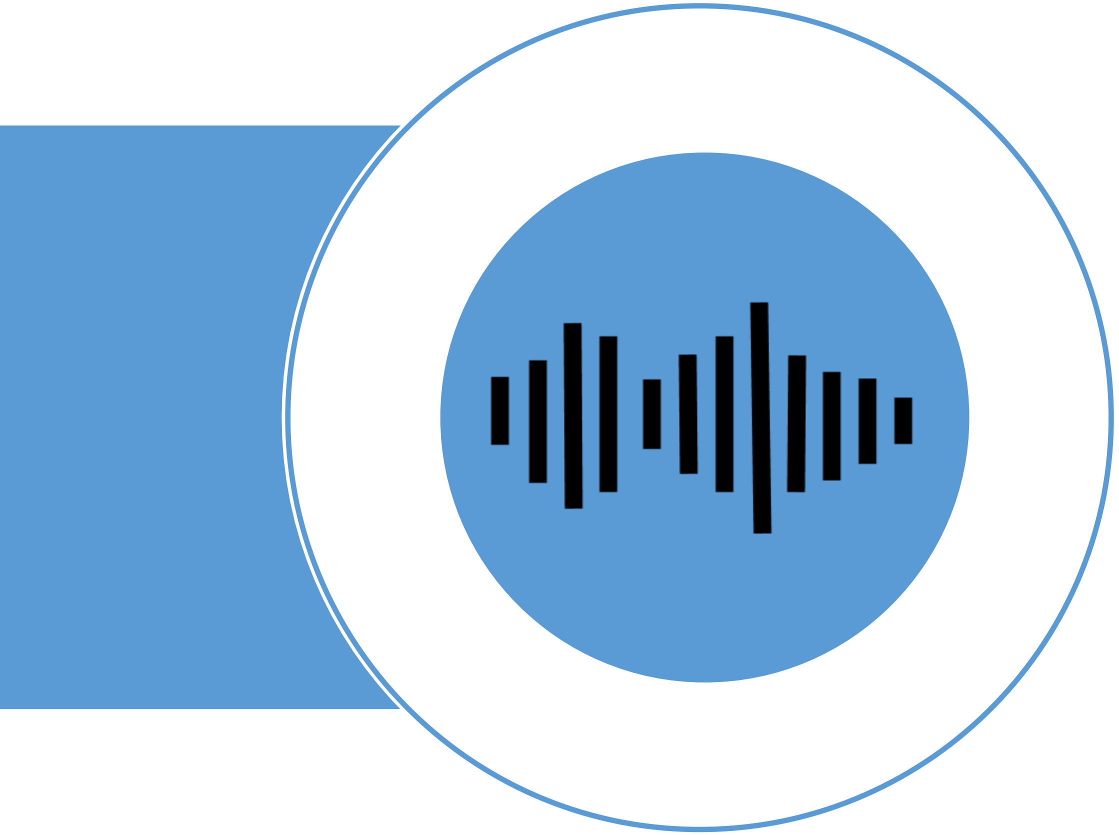
LJ speech 1.1 female speaker

AutoChild speech synthesis

AutoChild speech synthesis

- We build a new inference phase for AutoVocoder to synthesize high-quality child speech.
- can be used in real-time applications, low-latency and integration with other systems

AutoChild speech synthesis



AutoChild speech synthesis

1. **AutoVocoder Training**: Initially trained with single female adult LJ speech 1.1 "clean" dataset.

2. **Fine-Tuning for Child Speech**: Fine-tuning with pre-processed clean Hungarian child datasets.

3. **Training Details**: Utilized Adam optimizer, batch size of 16, and learning rate of 0.0002.

4. **Duration**: Training process extended over approximately three days, reaching 210,000 iterations.

AutoVocoder Training Setup

Model Training

- Batch size: 16
- Learning rate: 0.0002
- Training duration: 3 days
- Total epochs: 210,000

System Configuration

- System: Ubuntu 16.04.7 LTS
- GPU Resources: Utilized NVIDIA-SMI and CUDA version 11.4
- Efficient GPU utilization for faster training.

Modification for Child Speech

- Originally designed for adult speech synthesis.
- Adapted for child speech synthesis.

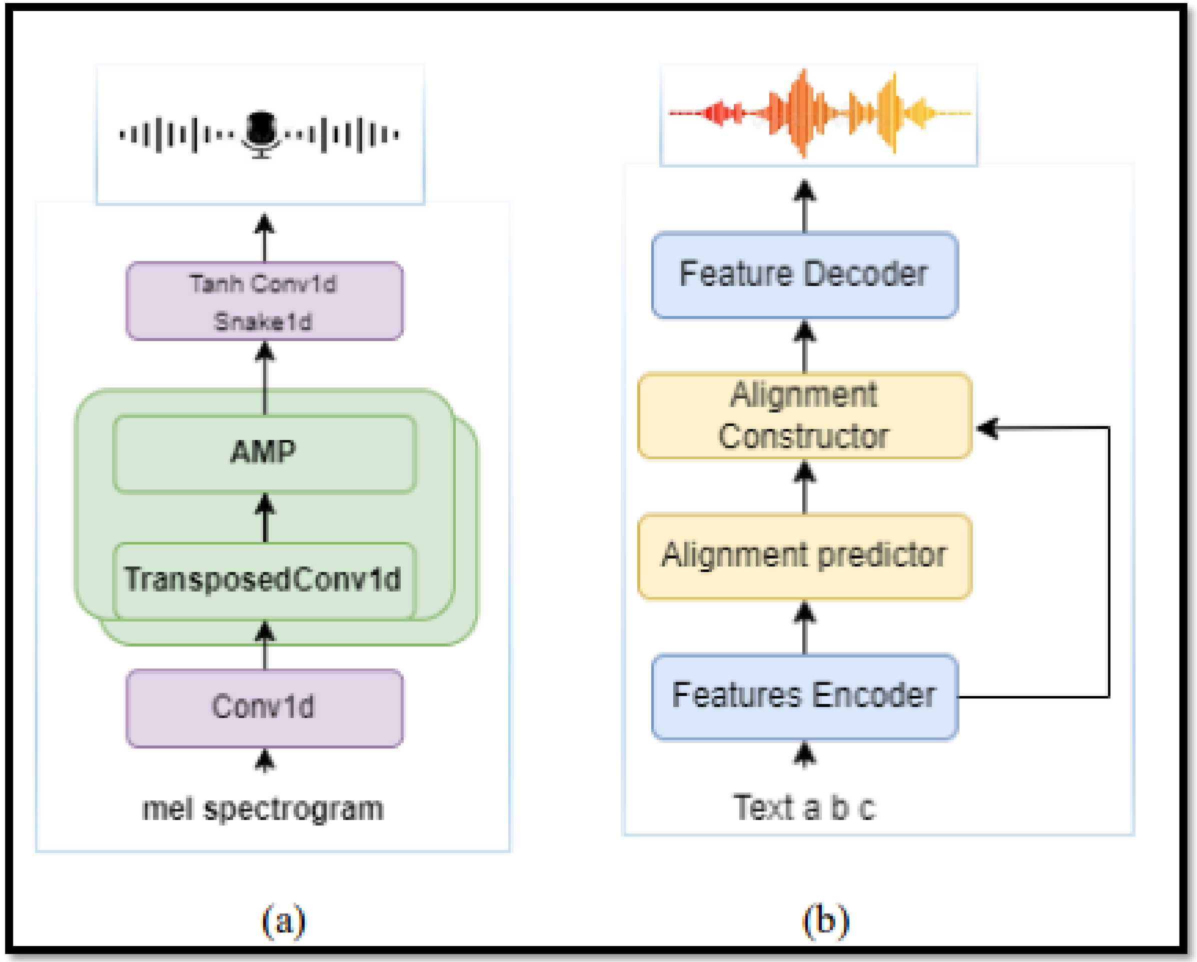
Baseline World and Parallel WaveGan

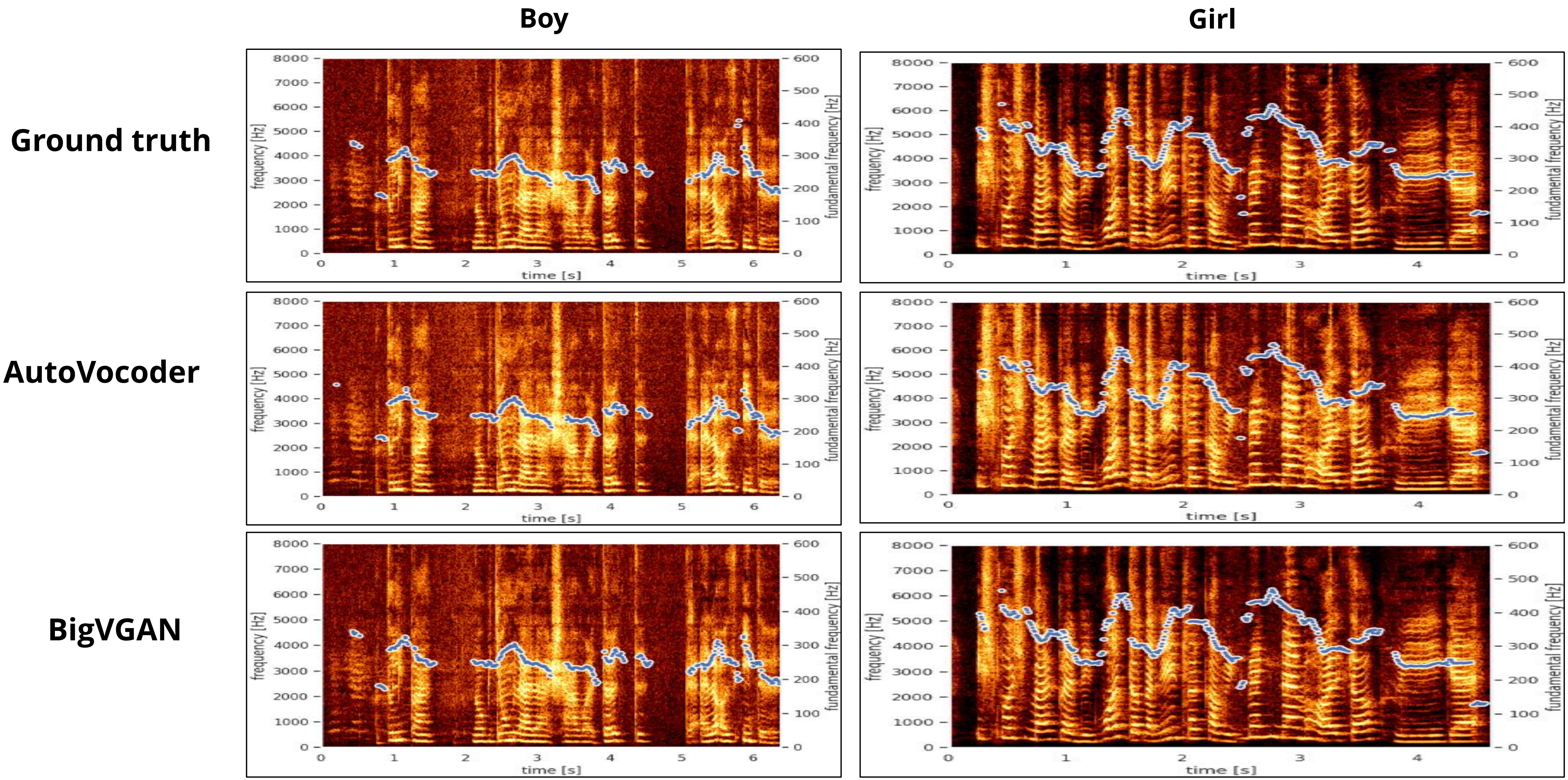
1. **BigVGAN Configuration**: employed in this study featured a Batch Size of 32, a Learning Rate of 0.0001, and underwent 500,000 iterations. with Adam optimizer.
2. **Comparison**: Evaluated AutoVocoder against BigVGAN in the quality of synthesized child speech.

Vocoders

AutoChild speech synthesis

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





Boy

Girl

Ground truth

AutoVocoder

BigVGAN

03

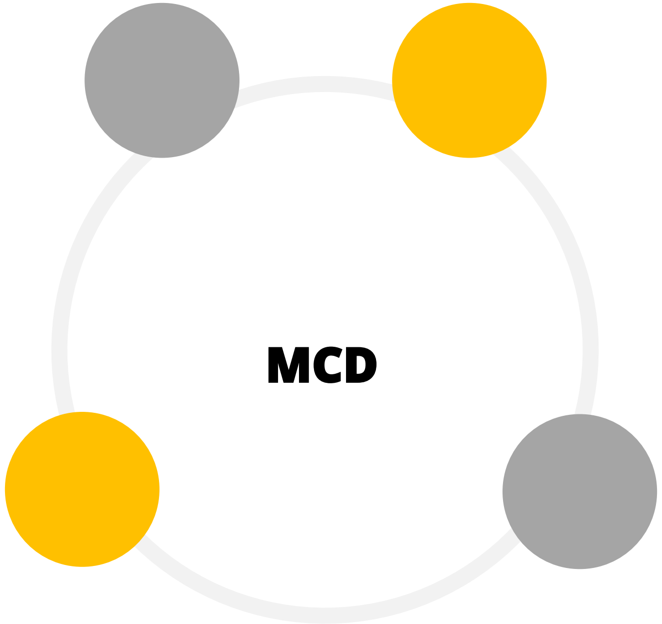
EXPERIMENTAL RESULTS

subjective and objective assessments

Mel-Cepstral Distortion MCD

Utilized MCD (Mel-Cepstral Distortion) as a distance metric for objective evaluation of synthesized sound quality.

Smaller MCD implies a higher similarity between synthesized and original speech.



Average MCD calculated for ten synthesized sound samples for both girls and boys for all models

MCD results indicate a high correlation between AutoVocoder and ground-truth data, suggesting closely matched characteristics.

Mel-Cepstral Distortion MCD

Systems	Boy	Girl
AutoVocoder	0.843	0.842
BigVGAN	**0.814**	**0.762**

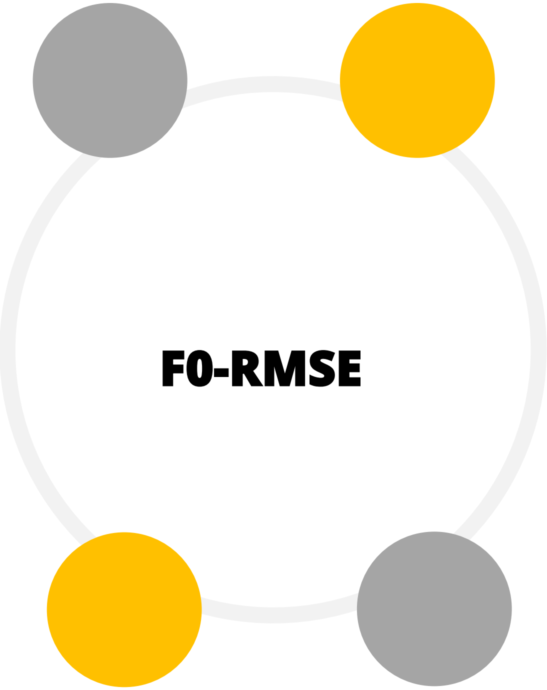
MCD

MCD

F0 Root Mean Square Error F0-RMSE

Employed F0 root mean square error (F0-RMSE) to compare log F0 values between ground truth and synthesized waveforms.

All evaluation metrics consistently highlight the AutoVocoder's exceptional performance



F0-RMSE

Smaller F0-RMSE indicates a smaller prediction error in log F0 values.

We noted that BigVGAN achieved the lowest F0-RMSE value for boys’ sounds. For Girl sounds, a similar order is observed regarding the F0-RMSE values

F0 Root Mean Square Error F0-RMSE

Systems	Boy	Girl
AutoVocoder	3.33	3.25
BigVGAN	**3.27**	**3.11**

F0-RMSE

F0-RMSE

Subjective evaluation

MUSHRA Listening Test

1. **MUSHRA Test**: Evaluation of two speech synthesis models: AutoVocoder, and BigVGAN vocoders.

2. **Conditions**: Participants assessed 12 samples under five conditions: (a) Hidden ground truth , (b) AutoVocoder, (c) BigVGAN, and (d) Low-pass attenuated at 3.5 kHz,

3. **Participants**: 10 listeners aged 22 to 65, 8 females and 16 males, with no hearing impairments, completed the test in an average of 15 minutes.

MUSHRA Test

Subjective evaluation

MUSHRA Test Results and Listener Preferences

1. **Performance Comparison:** The MUSHRA test results for speech synthesis models are presented.
2. **Boys vs. Girls**: In synthesized speech for boys, AutoVocoder closely matches the ground truth
3. **Girls' Speech**: AutoVocoder is on par with the ground truth, both scoring 80%,
4. BigVGAN, the model has more complex architecture with a multi-scale discriminator and requires more resources and longer training time due to the larger batch size compared to the AutoVocoder.
5. **Computational Efficiency**: Achieved with significantly lower computational cost, enabling real-time deployment on low-powered devices for child TTS systems.

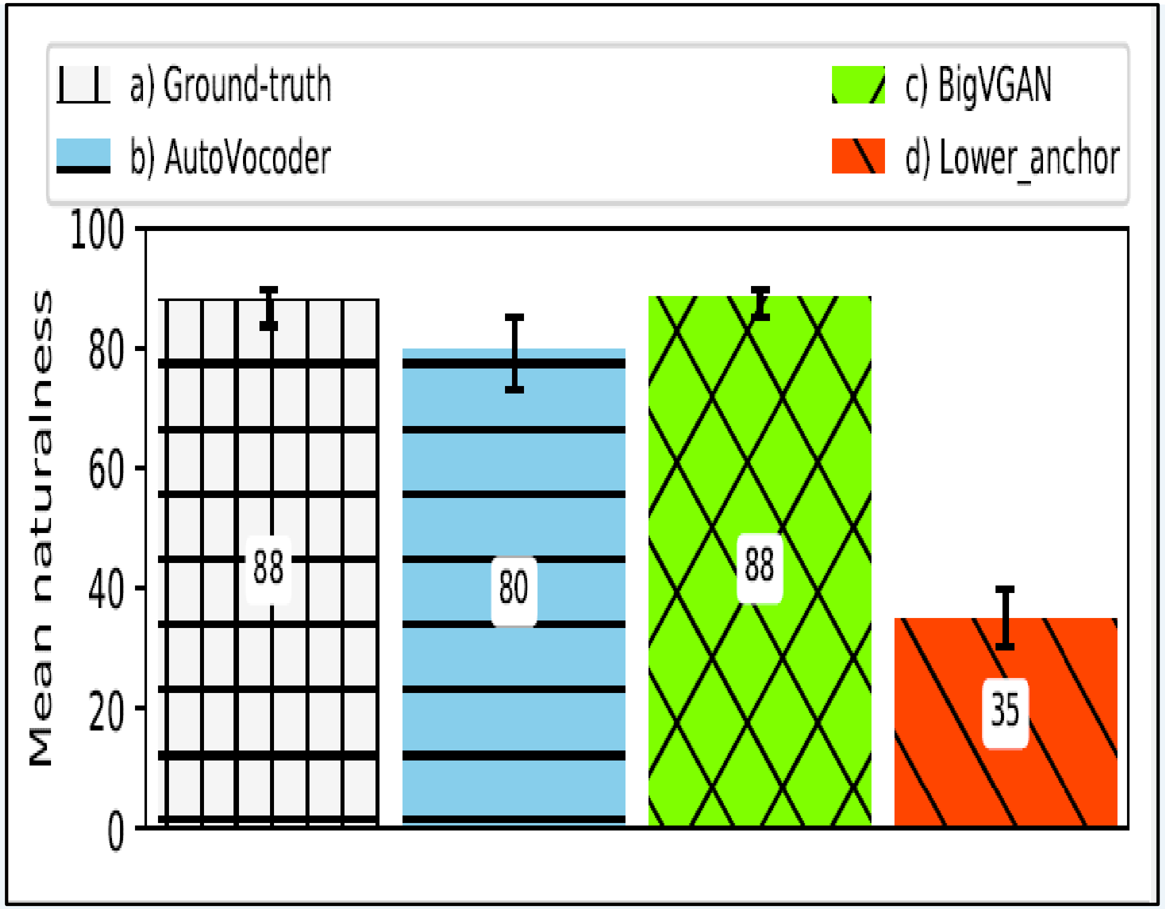
MUSHRA Test

Subjective evaluation

BOY	
Ground truth	🔊
BigVGAN	🔊
AutoVocoder	🔊

GIRL	
Ground truth	🔊
BigVGAN	🔊
AutoVocoder	🔊

The MUSHRA scores the average results



MUSHRA Test

1 AutoVocoder

Achieved 80% score Close to the hidden reference.

2 BigVGAN Vocoders

BigVGAN received higher scores for compared to the Autovocoder

3 Model Success

High score signifies its ability to match natural speech. Demonstrates the capability to produce highly natural speech.

CONCLUSION

CONCLUSION

- This work aimed to explore advanced speech synthesis models for multi child speech in Hungarian languages. Universal neural vocoders BigVGAN and AutoVocoder.
- The AutoVocoder, on the other hand, significantly enhanced the quality of synthesized child speech. The evaluation result showed that the original and synthesis speech is very close to each other.
- BigVGAN, the model has more complex architecture with a multi-scale discriminator and requires more resources and longer training time due to the larger batch size compared to the AutoVocoder.

Multi-Speaker Child Speech Synthesis in Low-Resource Hungarian Language

Thank you