ChildTinyTalks (CTT):
A Benchmark Dataset and Baseline for Expressive Child Synthesis

Shaimaa Alwaisi

Shaima.alwaisi@edu.bme.hu

Supervisors
Prof. Géza Németh
Dr. Mohammed Salah Al-Radhi

Nov. 27, 2024

Outline

- Challenges in Child Speech TTS
- ChildTinyTalk CTT
- Objective and Subjective Results
- Conclusion

Challenges in Child Speech TTS

Child speech characteristics

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

Data challenge

Child speech datasets (e.g., MyST, CMU Kids) suffer from issues like noise, transcription errors, and lack of expressive styles.

Dificullty in collecting data

Difficulty in collecting clean recordings due to articulation and environmental factors.

ChildTinyTalk CTT

Existing Datasets Overview

Examples:

- MyST, CMU Kids, OGI Kids, Tball, Providence.

Limitations:

- Noisy data.
- Lack of expressive styles.
- Misaligned transcriptions.

Contribution of CTT Dataset

Unique Features:

- High-quality expressive child speech.
- Four expressive styles: Sadness, Excitement, Happiness, **Dataset Size:**
- 1200 utterances. And 2 hours of speech.

Dataset Details

01

ChildTinyTalks CTT dataset

Speech collected from 25 kids in grades ranging from third to fourth grade 1200 audio samples (Two Hours)

02

Speech Styles

🙁 🙂 😐 👍

(sadness, happiness, neutral and excitement)

03

Description

Dataset	Statistics
Number of POI	**25**
Number of Utterances	**1200**
Number of hours	**2**
Number of filtered Videos	**100**
Number of videos per POI	**3**
Avg Number of utterances per POI	**48**
Avg length of utterances [s]	**6.71**

Methodology

1 Data Collection

- TEDx Talks for Kids on YouTube
- American English
- **25 children (10 boys, 15 girls), ages 6–11**
- Sadness, Excitement, Happiness, Neutral.

2 Audio Cleaning

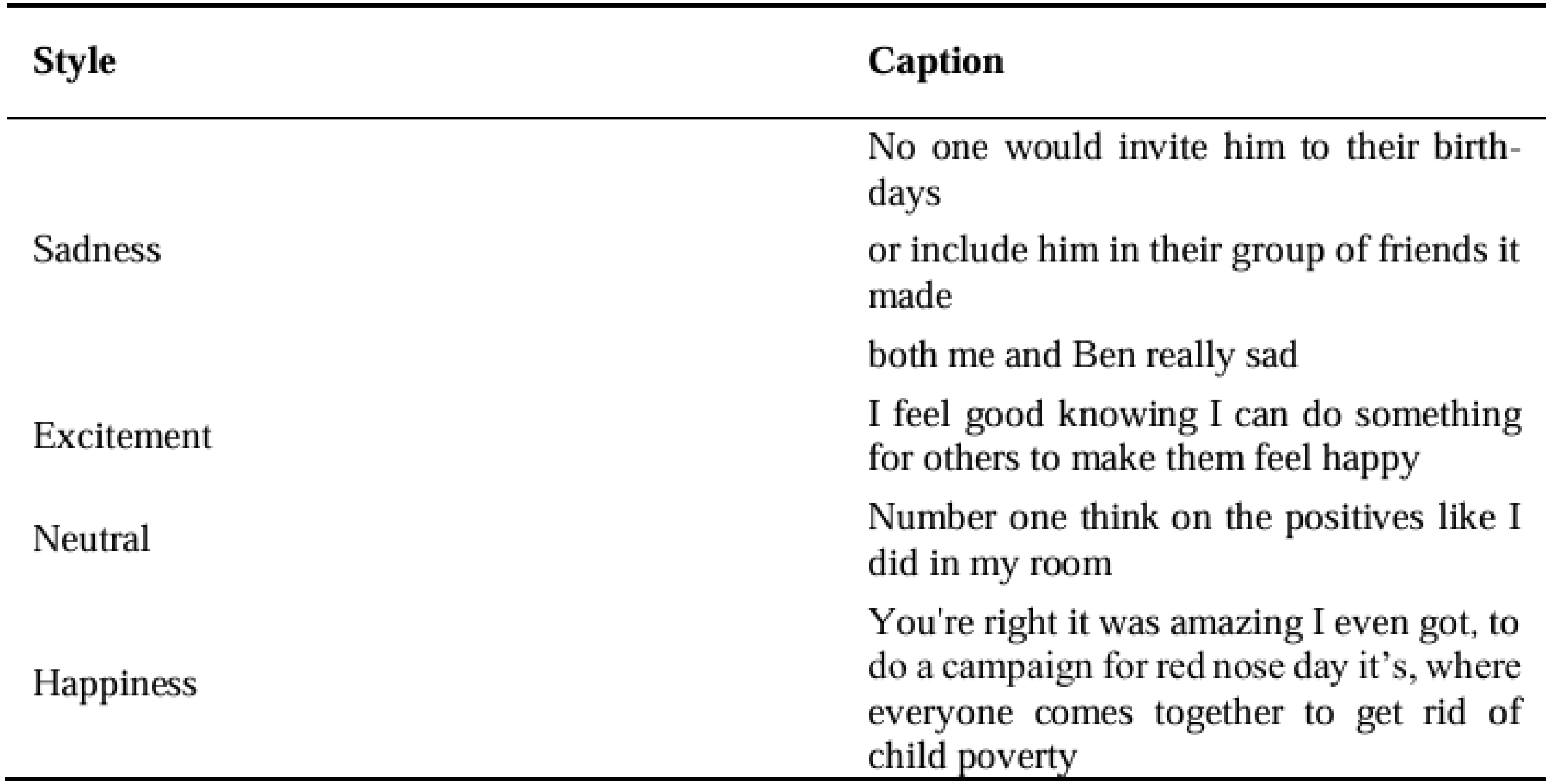
- **Audio Cleaning**:Removed background noise.
- Trimmed silences at the beginning and end.
- Transcription Refinement:Aligned text with audio at the word level.
- Resampling:Audio downsampled to 22.05 kHz for consistency.

3 Preprocessing

- All audio files decoded to WAV format.
- Removed substantial silence periods.
- Audio lengths averaged to 5.5 seconds.
- Short signals padded with zeros for consistency.

Examples of captions in ChildTinyTalk CTT

Table 2: Examples of captions in ChildTinyTalk CTT



Style	Caption
Sadness	No one would invite him to their birthdays or include him in their group of friends it made both me and Ben really sad
Excitement	I feel good knowing I can do something for others to make them feel happy
Neutral	Number one think on the positives like I did in my room
Happiness	You're right it was amazing I even got, to do a campaign for red nose day it's, where everyone comes together to get rid of child poverty

ChildTinyTalk CTT Description

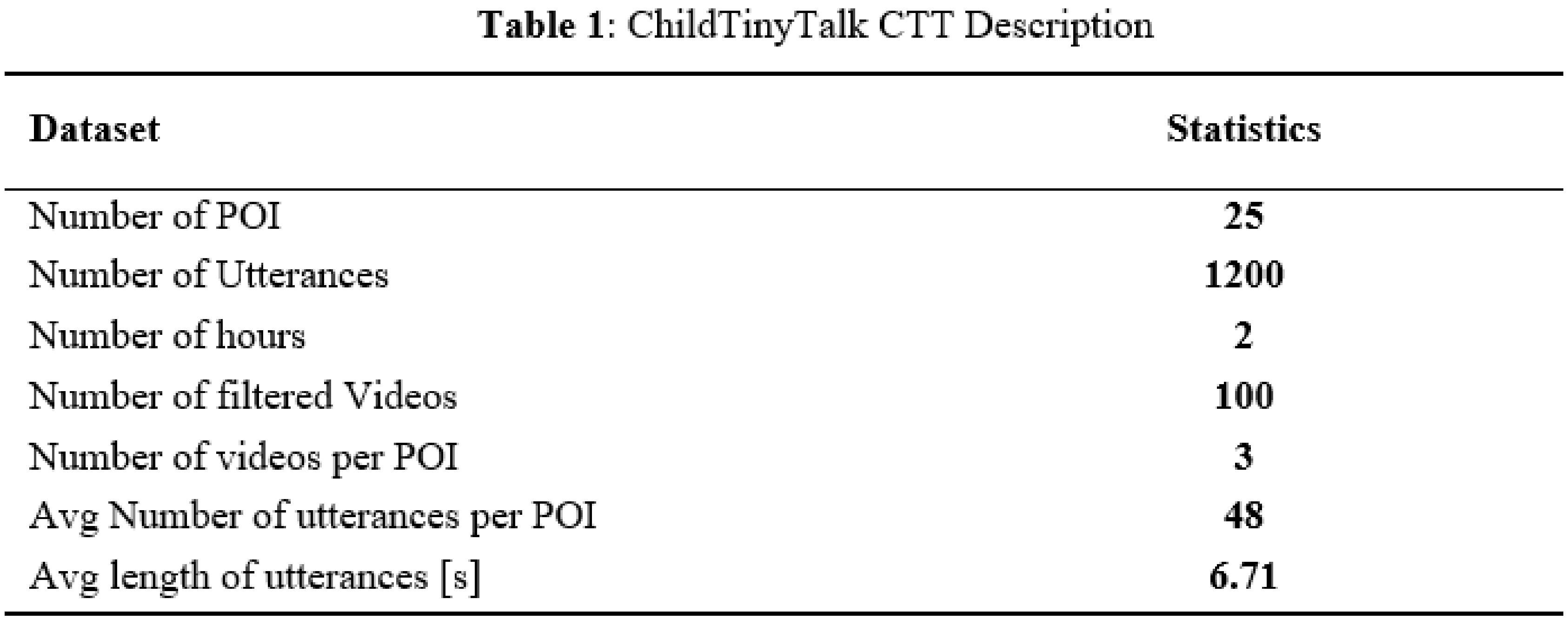


Table 1: ChildTinyTalk CTT Description

Dataset	**Statistics**
Number of POI	**25**
Number of Utterances	**1200**
Number of hours	**2**
Number of filtered Videos	**100**
Number of videos per POI	**3**
Avg Number of utterances per POI	**48**
Avg length of utterances [s]	**6.71**

CTT Efficiency

AutoVocoder Models:

- Trained on the **LJ Speech Dataset** and the custom **CTT Dataset**.
- Selected due to **state-of-the-art speech synthesis capabilities**.
- Excels in **fine-tuning child speech** .

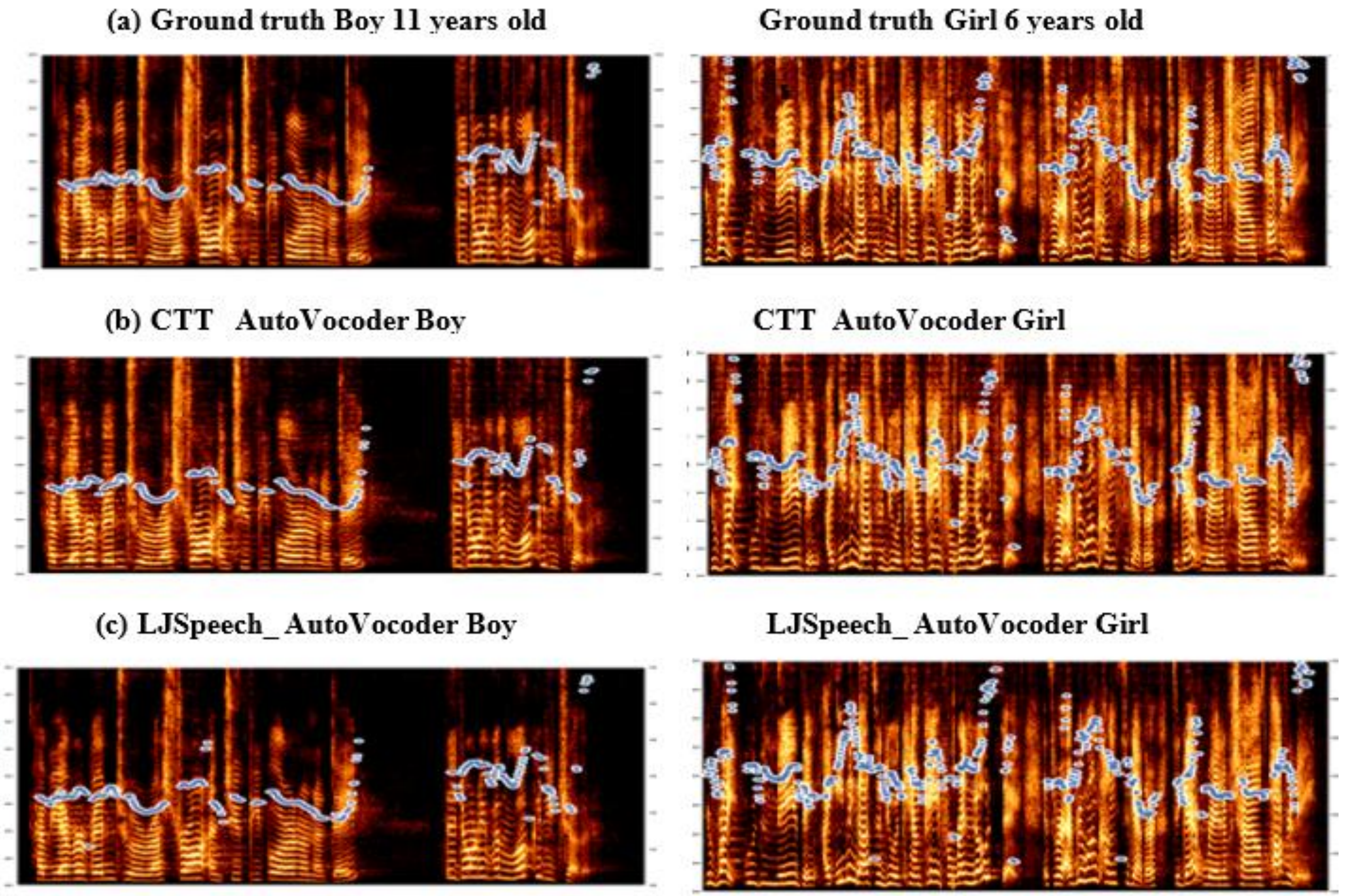
Outcome

•**Efficient Test Waveform Generation** using trained AutoVocoder models.

•Demonstrates the **practical efficiency of CTT** in real-world scenarios.

Mel Spectrogram

Fig 1. Example of melSpectrogram and F0 comparison between reference and child audio synthe-sized for a boy and a girl :(a) ground truth spectrogram and F0, (b) CTT_AutoVocoder spec-trogram and F0 Trained on our dataset CTT, and (c) LJSpeech_ AutoVocoder spectrogram and F0 trained on LJ speech 1.1 dataset. The horizontal axis gives the time dimension for the audio, while the left vertical axis represents the frequency dimensions. The right vertical axis represents the fundamental frequency



(a) Ground truth Boy 11 years old

Ground truth Girl 6 years old

(b) CTT AutoVocoder Boy

CTT AutoVocoder Girl

(c) LJSpeech_ AutoVocoder Boy

LJSpeech_ AutoVocoder Girl

HAVE YOU EVER BEEN TO A RESTAURANT AND TWO PEOPLE ARE ON A DATE THEY DON'T EVEN

YOU'RE RIGHT IT WAS AMAZINGI EVEN GOT, TO DO A CAMPAIGN FOR RED NOSE DAY IT'S, WHERE EVERYONE COMES TOGETHER TO GET RID OF CHILD

Objective Results

1. **Mel-Cepstral Distortion (MCD)**

   - Measures dissimilarity between two time-aligned mel-cepstral sequences.
   - Smaller MCD → Higher similarity between synthesized and original speech.
   - **Insights:** CTT_AutoVocoder performs comparably to LJ_AutoVocoder trained on adult speech.

2. **F0 Root Mean Square Error (F0-RMSE)**

   - **Definition**: Measures prediction error of fundamental frequency (F0).
   - Smaller RMSE → Lower prediction error.

Objective Results

Objective test

Systems	Boy	Girl
CTT_AutoVocoder	2.05	**1.84**
LJSpeech_AutoVocoder	**1.97**	2.13

MCD

MCD

Systems	Boy	Girl
CTT_AutoVocoder	2.96	3.23
LJSpeech_AutoVocoder	**3.00**	3.19

F0-RMSE

F0-RMSE

Subjective Results

•**MUSHRA (Multiple Stimuli with Hidden Reference and Anchor)**:

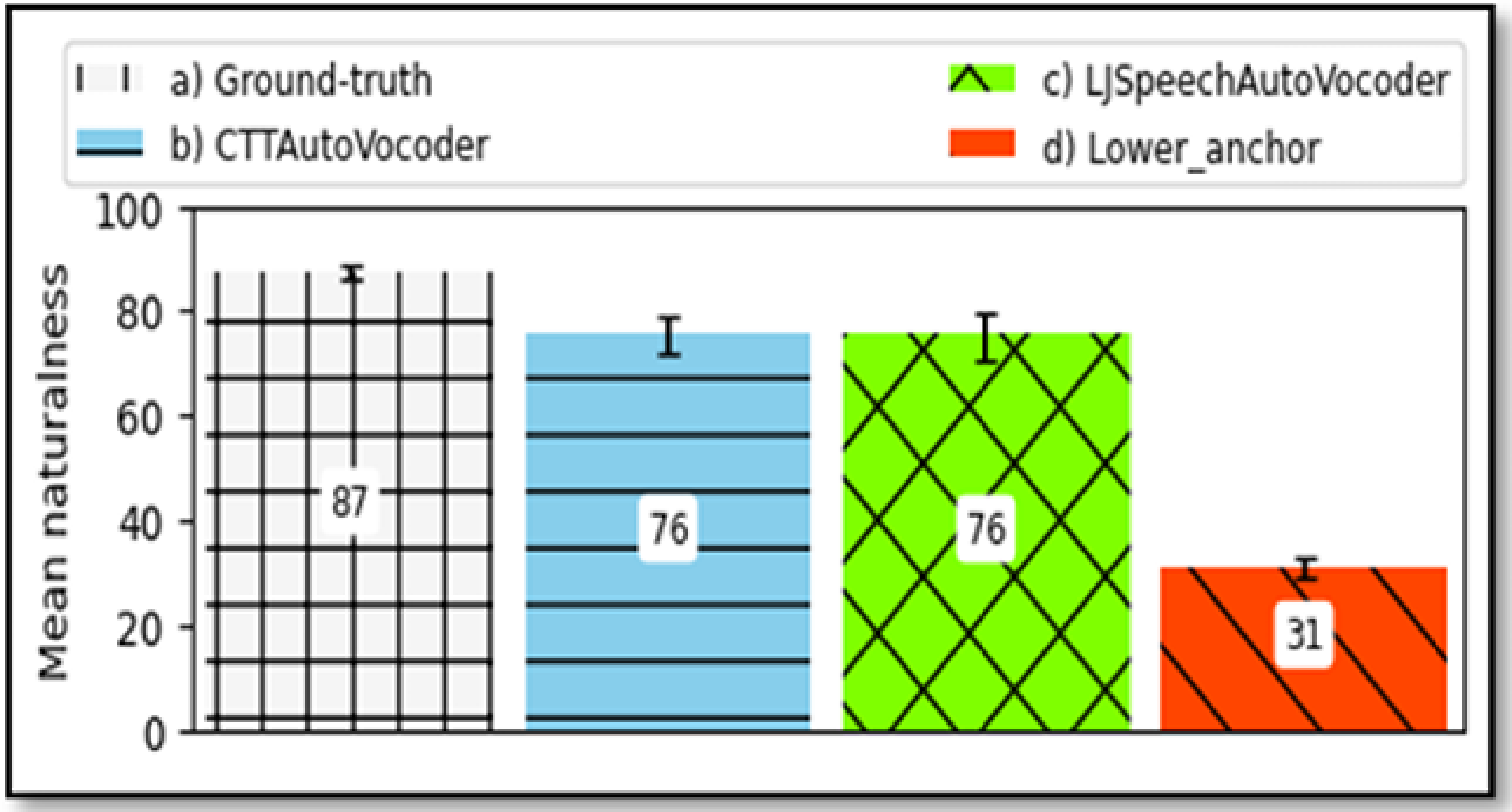
- **Purpose**: Compare the performance of AutoVocoder models.
- **Participants**: 23 listeners.
- **Task**: Rate the similarity of stimuli to the ground truth.
- Scoring Scale:Highly Not Similar → Highly Similar.

Results Summary

- **Mean Naturalness Scores**:
- **CTT_AutoVocoder** closely aligns with ground truth.
- **LJSpeech_AutoVocoder** shows comparable results.
- **76% of respondents** rated **CTT_AutoVocoder** as **very similar** to ground truth.

Subjective Results

Fig 2. The MUSHRA scores for the Mean naturalness are presented for (a) Ground truth (b) CTT_ AutoVocoder (c) LJSpeech_ AutoVocoder (d) Lower anchor, with the average results shown. A higher value indicates better overall quality.



Conclusion

Conclusion

- **Dataset**: 1200+ utterances, 25 children (ages 6–11), extracted from TEDx Kids Talks on YouTube.
- **Applications**: Child speech synthesis, TTS, and ASR systems.
- CTT dataset (~10% the size of large datasets) achieves **comparable results** to LJSpeech in acoustic modeling.
- **ASR Potential**: Helps address challenges in recognizing child speech.

Future Directions

- Dataset augmentation for size and expressiveness.
- Development of fully expressive child TTS synthesis.
- Adoption as a benchmark for child speech processing research.

ChildTinyTalks (CTT): A Benchmark Dataset and Baseline for Expressive Child Synthesis

Thank you