Children's Phonetic Speech Recognition

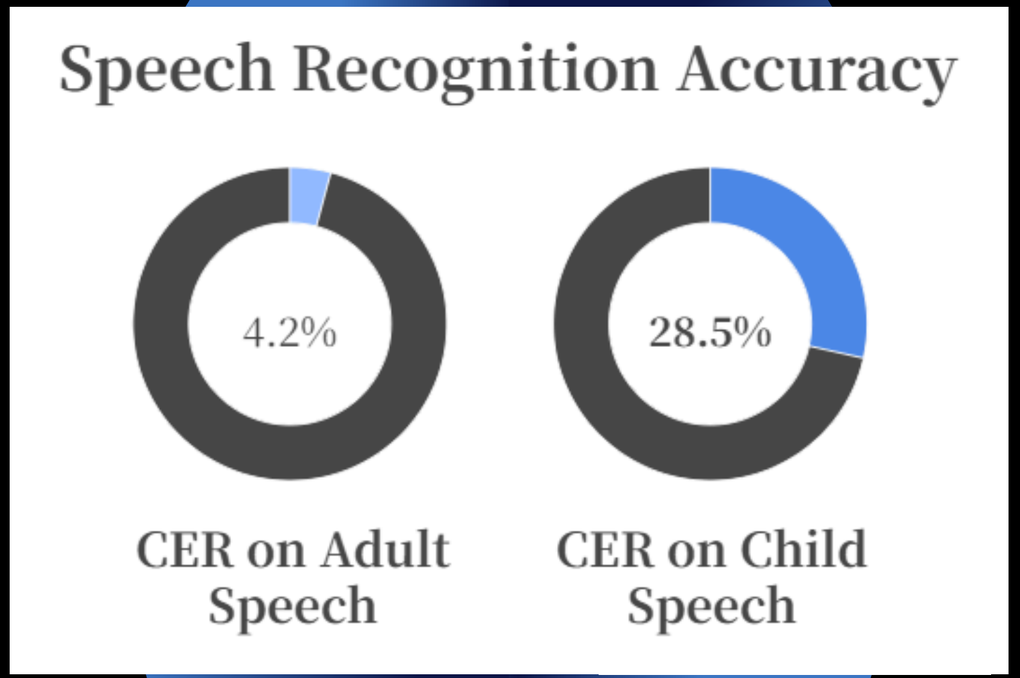
Armaan Raisinghani, Rohan Gupta, Shreya Khanna

MLPR End Semester Presentation

Problem Statement

Children’s speech recognition is significantly more difficult than adult ASR due to:

- Acoustic Variability: Higher pitch and shorter vocal tracts.
- Pronunciation: Emerging phonetic patterns and inconsistencies.
- Limited Data: Scarcity of large-scale annotated child corpora.



Applications and Impact

Enabling AI tutors and literacy tools tailored to younger voices.

AI tools can improve reading engagement by ~20–30% in early learners [1].

Automated screening for pediatric speech pathology.

1 in 12 children in the US have a speech disorder [2].

Closing the performance gap in consumer voice assistants.

Voice assistants respond appropriately to children only ~50% of the time [3].

Context

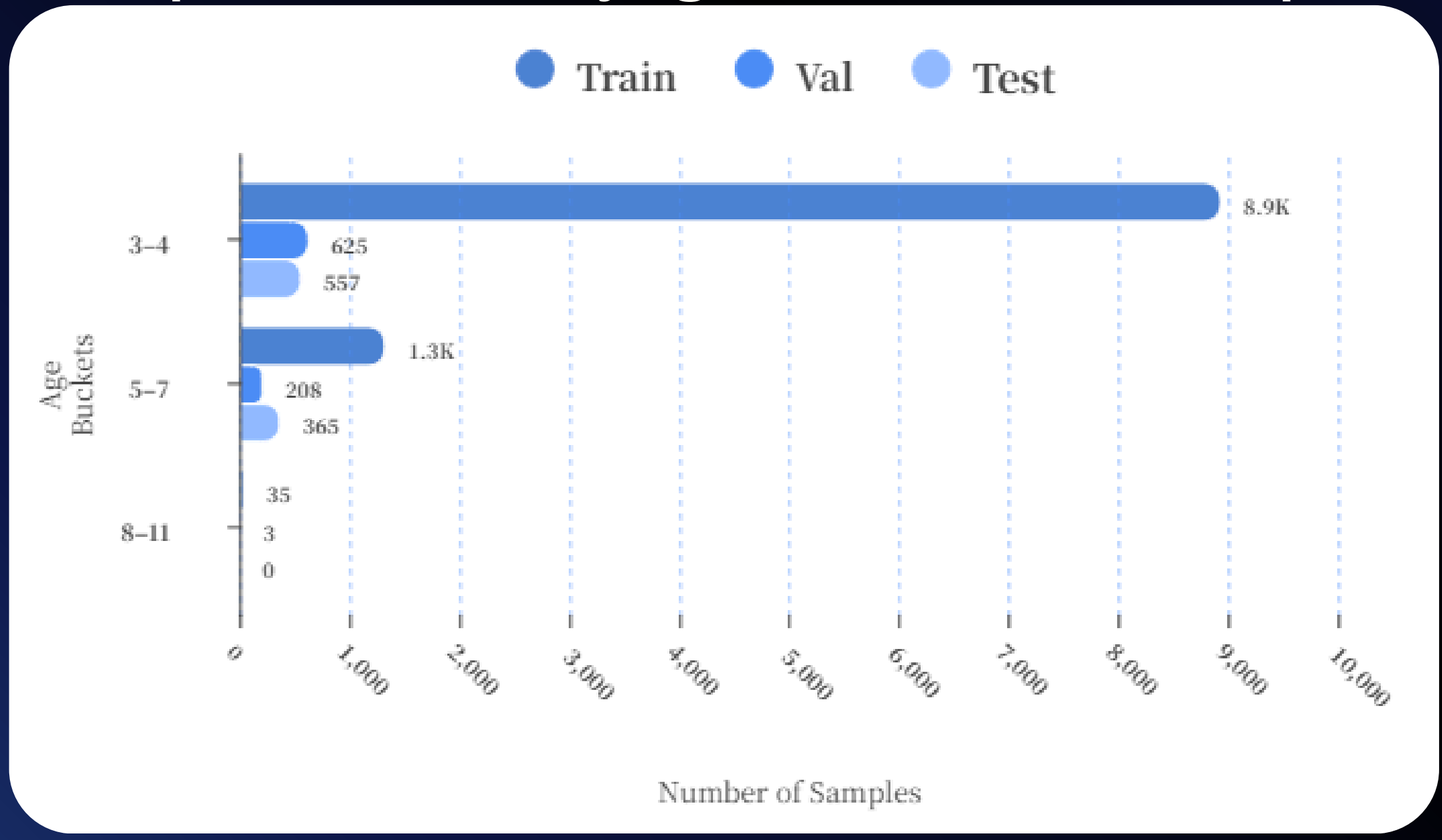
Goal: Build an ML model predicting phonetic symbols from children's speech audio.

Dataset: DrivenData Gates Foundation corpus. Audio was recorded from children reading aloud under supervised conditions. Ethical considerations: Data involves minors' voices. The dataset is provided through an official competition platform with institutional oversight. **Labels are IPA phonetic transcriptions aligned at utterance level.**

The metric computes the minimum number of **substitutions (S), deletions (D), and insertions (I)** required to transform the predicted character sequence into the reference sequence, divided by the **total number of reference characters (N):**

$$CER = \frac{S + D + I}{N}$$

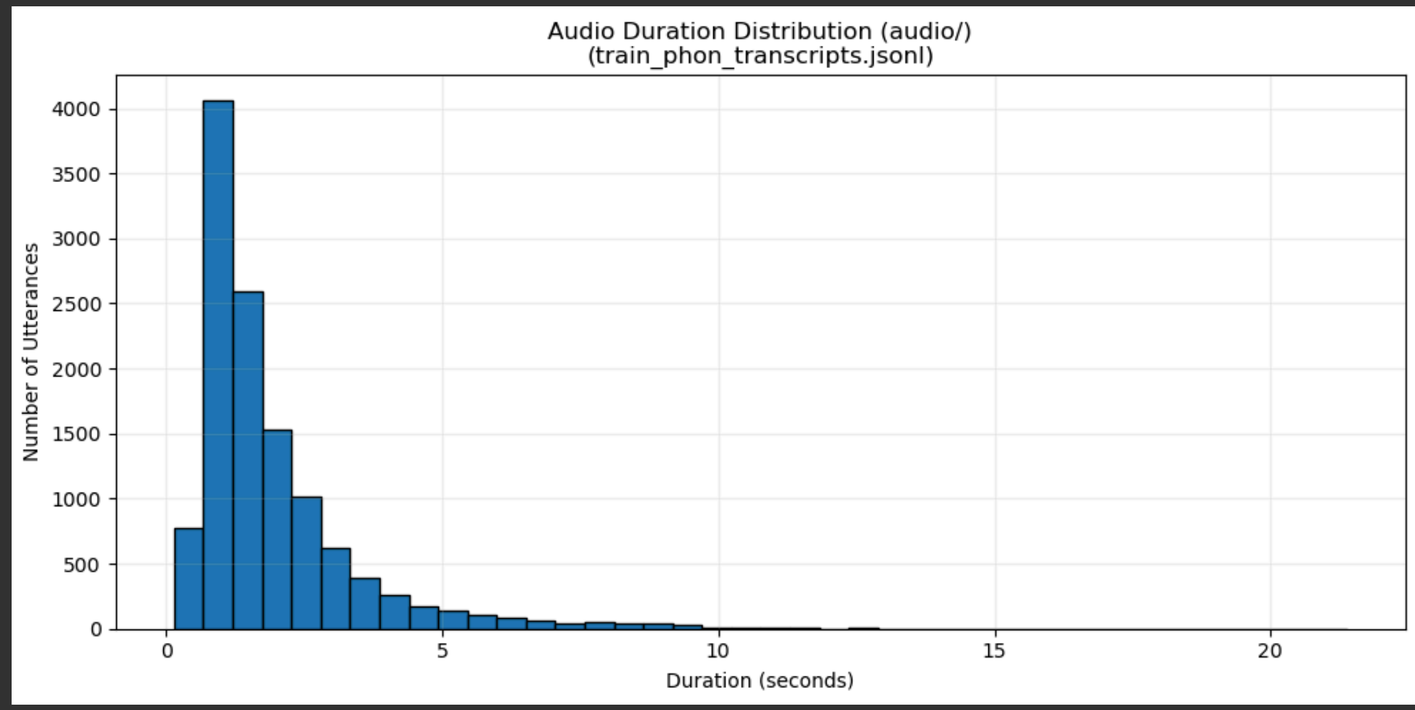
Sample Distribution by Age Bucket Across Data Splits

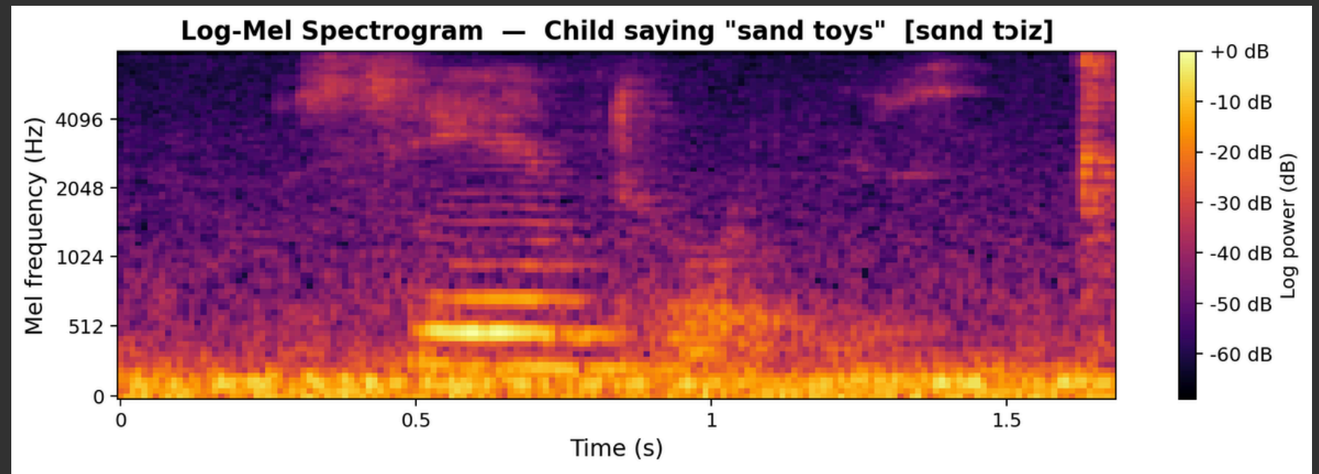


The Train split is heavily dominated by the 3-4 age bucket, while the Test split shows a more even distribution between 3-4 and 5-7 age groups.

Features Preprocessing & Extraction

- Audio loaded at 16kHz, amplitude normalised per sample
- Log-Mel spectrogram features extracted (80 mel bins) for CLDNN; MFCC + Δ + ΔΔ (80 features) for BiLSTM baselines
- IPA long vowels (ː) parsed with a custom tokeniser to avoid splitting two-character phones
- SpecAugment (time + frequency masking) applied after 3-epoch warmup to avoid destabilising early CTC alignment
- No PCA/LDA applied — the CNN frontend acts as implicit dimensionality reduction; frequency axis compressed 4× by strided convolutions while time axis preserved for CTC
- Variable-length clips handled by padding within batch + passing true lengths to CTC loss





Literature Review

CTC & BiLSTMs

Graves et al. (2006) [4]

- Introduced Connectionist Temporal Classification (CTC) for sequence labeling without frame-level alignment.
- Enabled training on unsegmented speech data.
- Combined effectively with Bidirectional LSTMs (BiLSTMs).
- Preserved temporal phoneme ordering across long sequences.
- Became foundational for modern end-to-end ASR systems.

DeepSpeech2

Amodei et al. (2016) [5]

- Proposed a scalable end-to-end ASR pipeline.
- Used deep recurrent architectures trained directly from audio.
- Reduced dependence on handcrafted speech features.
- Demonstrated strong performance on noisy and multilingual speech.
- Showed that large-scale training significantly improves ASR accuracy.

CLDNN Architecture

Sainath et al. (2015) [6]

- Combines convolutional, recurrent, and fully connected layers in a single ASR pipeline.
- CNN layers extract local acoustic patterns from speech signals.
- LSTM layers model temporal dependencies across audio sequences.
- DNN layers perform final feature discrimination and classification.
- Performs better than standalone CNN- or RNN-based architectures in speech recognition tasks.
- Improves robustness to speaker variation and background noise.

Adult-to-Child Transfer Learning

Shivakumar & Georgiou (2020) [7]

- Investigated transfer learning for children's speech recognition.
- Showed that adult ASR models can't be adapted to child speech effectively.
- Identified major challenges: acoustic variability, pronunciation differences, and limited child speech datasets
- Found that adapting both:
  - lower acoustic layers
  - higher linguistic/pronunciation layers
- improves recognition performance.
- Demonstrated transfer learning reduces data requirements for child ASR.

Early Methodology

Plain BiLSTM + CTC

- Convert raw audio → 80-dim log-Mel spectrogram (a time-frequency representation of sound)
- Feed spectrogram directly into a 2-layer Bidirectional LSTM — no feature extraction step
- LSTM reads the spectrogram forwards and backwards, learning the temporal order of phones
- CTC (Connectionist Temporal Classification) aligns the LSTM output to the target IPA phone sequence without needing frame-level labels

Why we tried it: Baseline to test if temporal modelling alone is enough, without any CNN for pattern extraction
Result: Surprisingly strong (Test CER 0.6315) — proved that capturing phone order is the most important factor

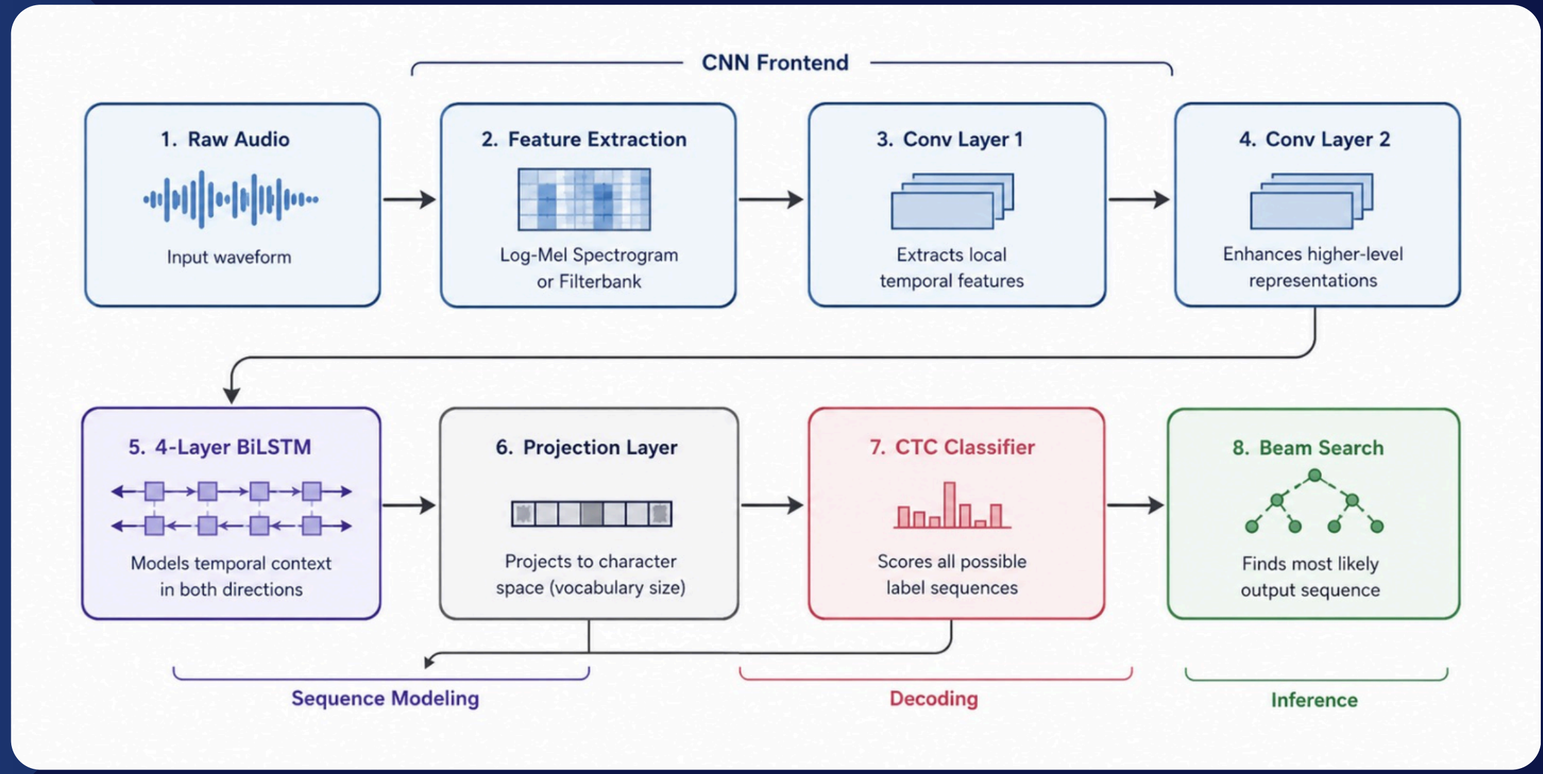
Spectogram + VGG + BiLSTM + CTC

- Convert raw audio → log-Mel spectrogram
- Treat the spectrogram as a grayscale image and pass it through a VGG-style CNN pretrained on ImageNet (1000-class natural image recognition)
- Use the CNN's extracted "visual" features as input to a BiLSTM for temporal modelling
- CTC aligns the output to IPA phones

Why we tried it: Spectrograms look like images — hypothesis was that pretrained image features (edges, textures) would capture spectral patterns (formants, harmonics)
Result: Severe overfitting (Test CER 0.886). Natural image features ≠ acoustic features. The domain gap was too large.

BEST RESULT- CLDNN: CNN-BiLSTM-CTC

- Convert raw audio → 80-dim log-Mel spectrogram
- Pass through a small 2-layer CNN trained from scratch on our data — learns local acoustic patterns (formant shapes, consonant bursts) directly from children's speech
- CNN compresses frequency while preserving full time resolution — critical for CTC, which needs enough time steps to emit all phones
- Feed CNN features into a 4-layer Bidirectional LSTM — captures long-range context (how phones flow into each other)
- Project through a bottleneck layer (768 → 256 dimensions) to reduce parameters
- CTC classifier outputs probabilities over 54 IPA phones + blank at each time step
- Beam search decoding finds the most likely phone sequence



- Why it worked: ASR-native design. The CNN learns acoustic features appropriate for children's speech (not borrowed from ImageNet or adult speech). The BiLSTM handles sequencing. The architecture is right-sized (17M params) for our 5.3 hours of data.
- Informed by: Sainath et al. (2015) CLDNN, Amodei et al. (2016) Deep Speech 2

Failed Experiments

Wav2Vec2 Feature Extraction

- Feed raw 16kHz waveform directly into a 7-layer 1D convolutional network (same architecture as the wav2vec2 feature extractor, but randomly initialised — no pretraining)
- The CNN is supposed to learn its own acoustic features from raw audio, replacing handcrafted log-Mel spectrograms
- CNN output feeds into the same 4-layer BiLSTM + CTC pipeline as the CLDNN

- *Why we tried it: Test whether end-to-end learned features can outperform handcrafted spectrograms*
- *Result: Val CER 0.811 — much worse than log-Mel CLDNN (0.651). The wav2vec2 CNN architecture was designed for 60,000+ hours of self-supervised pretraining. With only 5.3 hours, it cannot learn meaningful acoustic representations from scratch. Log-Mel spectrograms give a strong prior for free.*

DeepSpeech2 Transfer Learning

- Start with a full DeepSpeech2 model pretrained on 1000 hours of adult English speech (LibriSpeech)
- Freeze the entire pretrained encoder (CNN + 5-layer LSTM, 86.6M params) — do not update these weights
- Replace only the output layer with a new IPA phone classifier and train that on children's data

- *Why we tried it: If adult speech representations are general enough, just swapping the output head should adapt the model to children's IPA phones — similar to how ImageNet transfer works in vision*
- *Result: Val CER 0.798, only 44% of expected phones emitted. Adult-child acoustic mismatch (higher pitch, shorter vocal tracts, different pronunciation patterns) affects every layer of the encoder, not just the output. Freezing doesn't work — validates Shivakumar & Georgiou (2020).*

DeepSpeech2 From Scratch

- Same full DeepSpeech2 architecture (2-layer CNN + 5-layer LSTM, 86.6M params), but all weights randomly initialised — no pretrained checkpoint
- Trained end-to-end on children's data with SortaGrad (length-sorted first epoch for stability) and SpecAugment

- *Why we tried it: Since frozen transfer failed due to adult-child mismatch, does training the full DS2 architecture from scratch on children's data beat our smaller CLDNN?*
- *Result: Val CER 0.666, Test CER 0.619 — never reached CLDNN's 0.560. Over-parameterised: 86M parameters for 5.3 hours of data (vs CLDNN's 17M). The right model is the one sized for the data you have.*

Error Analysis

Audio as spectrogram

VGG-ImageNet features failed because spectrograms lack the spatial structure (edges, objects) that image models are trained on. We tried it because we knew transformers can handle this, but it failed on classical ML.

Adult → Child transfer

The research after the failed attempt showed that last layer tuning is not enough for adapting an adult speech model for children.

Deletions

Deletions are the dominant error mode: Best model emits only 72% of expected phones (8,123 vs 11,258) — the model tends to skip phones rather than hallucinate new ones.

Confusing confusions

The confusions made by the model are very similar in sounds (so much in fact that most languages/dialects do not even differentiate them), and thus there are no drastic misclassifications.

Summary

Stage	Model	Val CER	Test CER	Test PER	Main Finding
1	Plain BiLSTM	0.762	0.632	0.635	Temporal modelling was very important
2	VGG + BiLSTM	0.901	0.886	0.889	Spectrogram-as-image transfer overfit badly
3	CLDNN h256	0.655	0.565	0.579	Smaller ASR-native model was competitive
4	CLDNN h384	0.651	0.561	0.581	Best overall CER
5	Wav2Vec2 1D CNN (scratch)	0.811	0.759	0.800	Architecture without pretraining worse than log-Mel; insufficient data
6	DS2 frozen → IPA head	0.798	0.689	0.690	Adult frozen features fail on children's IPA; mismattch confirmed at all layers
7	DS2 from scratch	0.666	0.619	0.632	Larger DS2 worse than CLDNN; over-parameterised for 5.3h

References

[1] Frontiers in Education. “Artificial Intelligence and Reading Engagement in Early Learners.” Frontiers in Education, 2024.

[2] National Institute on Deafness and Other Communication Disorders (NIDCD). “Statistics on Voice, Speech, and Language.” National Institutes of Health (NIH). https://www.nidcd.nih.gov/health/statistics/statistics-voice-speech-and-language

[3] International Journal of Child-Computer Interaction. “How Commercial Voice Assistants Respond to Children’s Speech.” ScienceDirect, 2022. https://www.sciencedirect.com/science/article/abs/pii/S2212868922000587

[4] Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. “Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks.” Proceedings of the 23rd International Conference on Machine Learning (ICML), 2006.

[5] Amodei, D., Ananthanarayanan, S., Anubhai, R., et al. “Deep Speech 2: End-to-End Speech Recognition in English and Mandarin.” International Conference on Machine Learning (ICML), 2016.

[6] Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. “Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks.” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.

[7] Shivakumar, P. G., & Georgiou, P. “Transfer Learning from Adult to Children for Speech Recognition: Evaluation, Analysis and Recommendations.” Computer Speech & Language, 2020.

fin.