AI-Based Reading Assistant for Children Using Wav2Vec2

SCS_3546_038 Deep Learning
Term Project
Nataliia Kobrii

Problem Overview

- Early readers need support with pronunciation and phonics.
- Human feedback is not always available during home reading.
- Children’s speech is challenging for ASR (higher pitch, shorter phonemes, inconsistent articulation).

Goal: evaluate whether a pretrained speech model can provide useful, child-friendly pronunciation feedback.

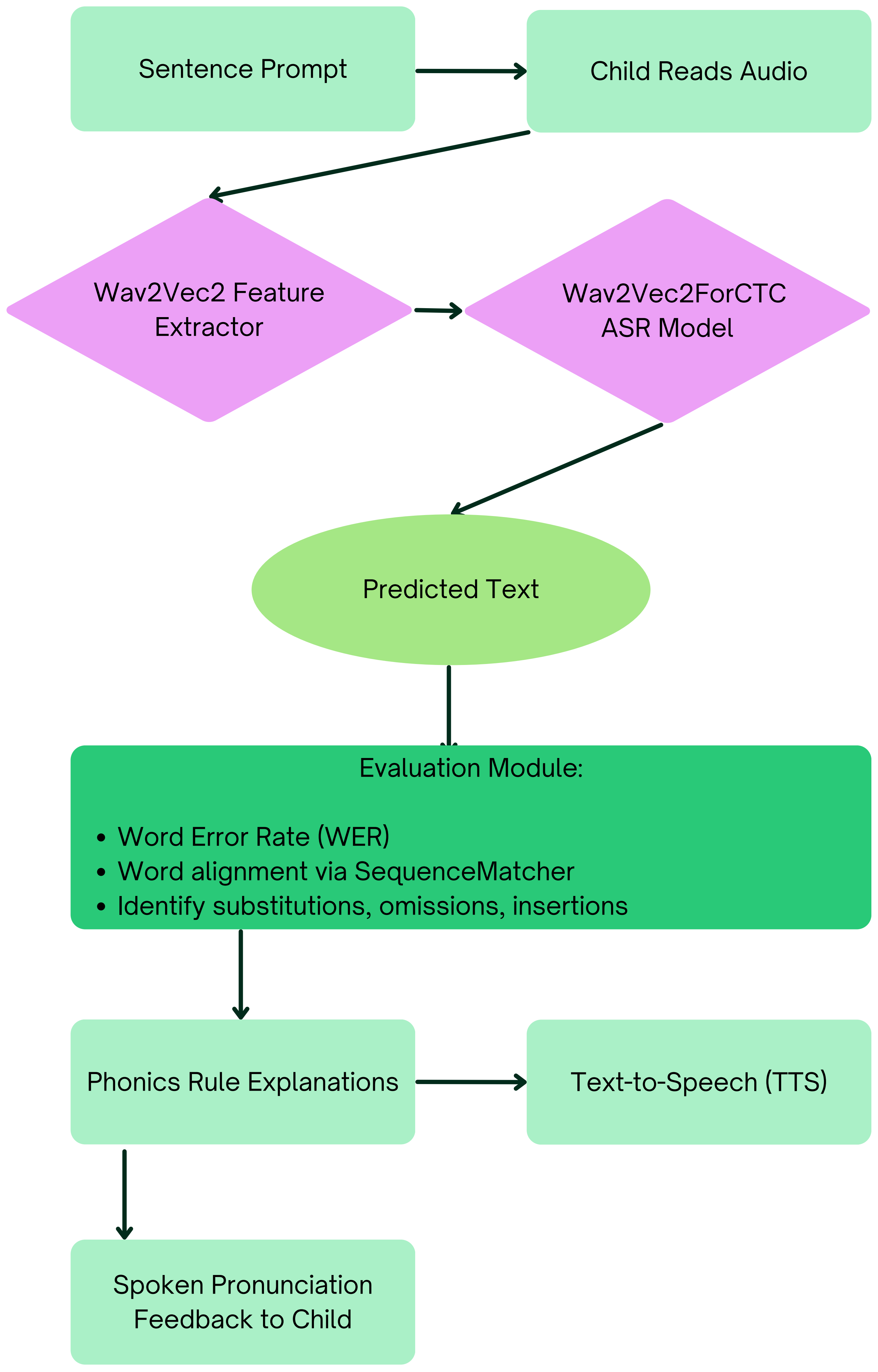
System Overview & Methods

Pipeline:

- Child reads a short sentence (10 custom sentences).
- Audio is transcribed using Wav2Vec2-Base-960h.
- Predicted text is compared to the target using WER (Word Error Rate).
- Word alignment detects incorrect or missing words.
- Phonics rules generate simple explanations for each error.
- Text-to-speech provides spoken feedback.

Methods:

- Model: Wav2Vec2ForCTC (pretrained on adult speech).
- Metrics: WER + pronunciation score = 1 - WER.
- Alignment: Python SequenceMatcher.
- Rules: “th”, short “a”, long “ee”, magic “e”.


Sentence Prompt
Child Reads Audio
Wav2Vec2 Feature Extractor
Wav2Vec2ForCTC ASR Model
Predicted Text
Evaluation Module:
• Word Error Rate (WER)
• Word alignment via SequenceMatcher
• Identify substitutions, omissions, insertions
Phonics Rule Explanations
Text-to-Speech (TTS)
Spoken Pronunciation Feedback to Child

Results From 10 Child Reading Samples

- **Average WER: 0.62** - transcription accuracy was limited.
- **Best result:** Sentence s08 (Score 0.75).
- **Hardest sentence:** s03 (Score 0.00).

- **Model consistently struggled with:**

1. "th" (e.g., "the" -> "sa/zasan/that")
2. vowel contrasts ("cat" -> "ket")
3. ending sounds ("sleeping" -> "slip")

Model performed well when pronunciation was clear and slow.

SENTENCE	WER	SCORE
Best (s08)	0.25	0.75
Worst (s03)	1.00	0.00
Average	0.62	0.38

Error Analysis & Strengths

ASR Limitations:

- adult-trained model -> poor child generalization
- common confusions: “th”, vowels, r-clusters, “ing” endings

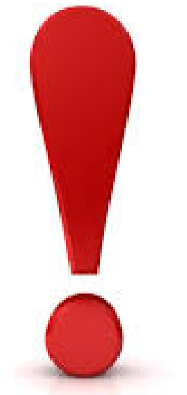
Strengths of the Prototype:

- Automatically identifies mispronounced words
- Generates rule-based phonics explanations
- TTS provides accessible audio guidance
- Works as a functional early-literacy support tool

Emergent patterns:

- Model tends to “collapse” multiple unclear words into one guess (“zasan”).
- Clear articulation leads to high accuracy.

Conclusion & Future work



- Pretrained **Wav2Vec2 can assist children’s reading**, despite limited accuracy.
- **Phonics-rule explanations compensate for transcription errors.**
- Prototype demonstrates educational potential.

Future improvements:

- Fine-tune Wav2Vec2 on child speech
- Add noise reduction + silence trimming
- Expand phonics rule library
- Develop an interactive UI
- Collect a larger child-speech dataset

Thank you!