Enhancing Computer-Assisted Pronunciation Training (CAPT) with Hybrid and End-to-End Children ASR Models

Aditya Kamlesh Parikh *(aditya.parikh@ru.nl)*, Cristian Tejedor-García, Catia Cucchiarini, Helmer Strik
Center For Language Studies (CLS), Radboud University Nijmegen, The Netherlands

Introduction

- ❏ **Mispronunciation Detection:** A key component of CAPT systems, that helps to detect the pronunciation errors in speech.
- ❏ **Confidence-Based Pronunciation Error Detection**: Uses ASR models to assess pronunciation via confidence scores at the frame/phoneme level, following forced alignment.
- ❏ While there is an established method for hybrid ASR models for calculating pronunciation scores, there are **no such standard approaches** to calculate pronunciation scores for **End-to-end (E2E) ASR models**.
- ❏ We utilized **entropy** and **logit margin** based pronunciation assessment scores **with E2E ASR models** and compared it with hybrid ASR's pronunciation assessment scores.

Research Questions

- **RQ1**: Can self-supervised (SSL) **E2E** pretrained models (e.g., XLS-R) fine tuned with native children's speech effectively recognize pronunciations and provide insights into phonetic and phonological properties compared to statistical hybrid ASR models?
- **RQ2**: To what extent our novel confidence measures for E2E models correlate with the model hypothesis?
- **RQ2.1:** Do they work better than hybrid ASR confidence scores?

Dataset

- **Jasmin-CGN Corpus** (Speech corpus from **Dutch** and Flemish children, non-native children, and elderly).
- Training data: **11,529 utterances from native Dutch and Flemish children** aged 7-11 (7,345 utterances) and 12-16 (4,184 utterances). From all utterances: 7,285 **read** speech utterances and 4,244 **dialogues**.
- Testing data: **6687 utterances from non-native children**, aged 7-16 years.

Methods

- **Trained hybrid Dutch ASR** based on **Time-delayed Neural Network (TDNN)** and phoneme based **statistical language models**.
- Calculated confidence scores at each pronunciation based on **Minimum Bayes Risk (MBR)**
- Fine Tuned end-to-end **(E2E)** SSL **XLS-R** model 300 million parameters.
- Calculated **confidence measures** with **Logit Margin** and advanced **entropy** based scores.
- Forced alignment algorithm: **Needleman-Wunsch**

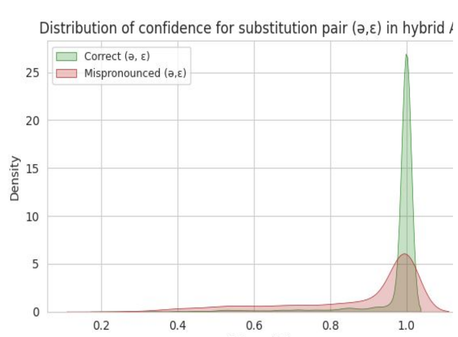
Results

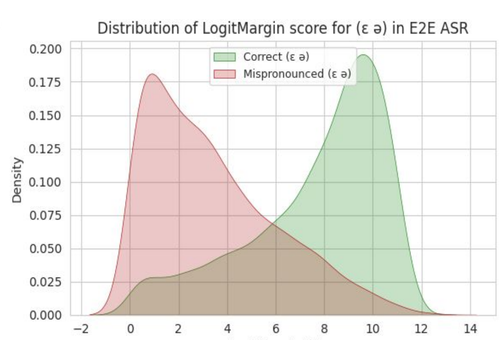
Phoneme Error Rate (PER)
Average children speech: 23.1% [Gao et al . 2024]
Non-native children speech:
- TDNN Hybrid ASR model: **12.53%**
- E2E XLS-R ASR model : **11.82%**

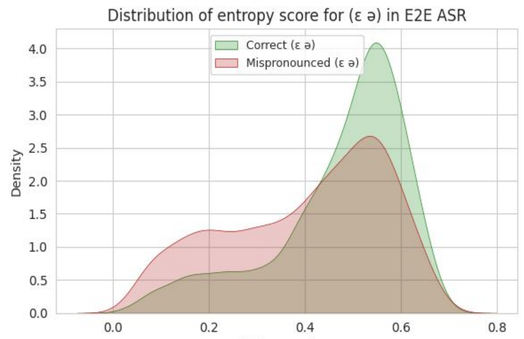
	Hybrid ASR confidence score	E2E ASR logit margin scores	E2E ASR entropy scores
Accuracy	**79.09%**	70.82%	69.30%
Precision	92.12%	**97.55%**	93.45%
Recall	**82.96%**	79.82%	71.25%
F1	87.30%	**87.80%**	80.85%

Most confusing pronunciations pairs				Most Insertions		Most Deletions	
Hybrid ASR		E2E ASR		Hybrid ASR	E2E ASR	Hybrid ASR	E2E ASR
ɑ	aː	ɑ	aː	ə	ə	ə	ə
ɛ	ə	ɛ	ə	n	n	t	n
d	t	t	d	t	t	ɛ	t
oː	ɔ	ɪ	ə	r	r	n	ɛ
ɑ	ɛ	m	n	h	d	r	r

Distribution of pronunciation scores when ɛ mispronounced as ə and vice versa







Discussion and Conclusion

Ans RQ1: Yes, the E2E XLS-R model outperforms the TDNN Hybrid ASR model, albeit by a small margin.
Ans RQ2: The focus is on the confidence measures for each phoneme, where the **Logit Margin scores** in the **E2E** ASR show **superior** performance, achieving an **F1 score of 87.80%** and **better distribution** for pronunciation scores. Logit Margin provides a clearer view of the model's output before the softmax compresses values into probabilities.
- E2E models with Logit Margin based pronunciation scores can be effective for pronunciation assessment.

Responsible AI for Voice Diagnostics
AiNed Fellowship Grants
NGF.1607.22.013 (RAIVD project)
AiNed
Radboud Universiteit