2024「中技社科技獎學金」

2024CTCI Foundation Science and Technology Scholarship

研究獎學金

Research Scholarship

An Effective Pronunciation Assessment Approach Leveraging Hierarchical Transformers and Pre-training Strategies

師大 | SMIL

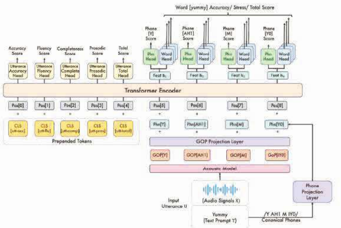
Bi-Cheng Yan, Jiun-Ting Li, Yi-Cheng Wang, Hsin-Wei Wang, Tien-Hong Lo, Berlin Chen

1 Highlights of This Work

- Automatic pronunciation assessment (APA) aims to quantify oral proficiency and provide multiple aspect scores at various linguistic granularities to language learners
  - Aspects: Accuracy, Fluency, Stress, etc.
  - Granularities: Utterance, Word, and Phone
- Our Contributions
  - HierTFR is proposed for hierarchically representing an input utterance and capturing relatedness within and across different linguistic levels
  - A correlation-aware regularizer is used for model training, which encourages prediction scores to consider the relatedness among disparate aspects
  - Extensive sets of experiments carried out on a public APA dataset confirm the utility of our proposed pre-training strategies

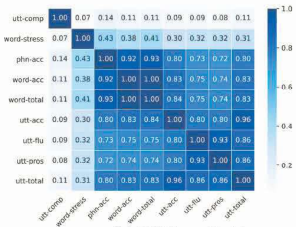
2 Motivations: Limitations of Parallel Modeling

- These methods fall short in taking advantage of the hierarchical structure of an utterance, assuming all phones within a word are of equal importance and insufficiently capturing the word-level structure cues in an utterance
- Second, most of these methods largely overlook the relatedness among the pronunciation aspects



- (a) GOPT

- An iconic APA model with parallel neural modeling



- (b) PCC Score Matrix

- Each element in the matrix corresponds to the PCC score of a pair of measured aspects

- Optimization
  - The overall loss can be expressed by

$$\mathcal{L} = \mathcal{L}_{APA} + \lambda \mathcal{L}_{cor},$$

- $\lambda \in [0,1]$ is a tunable parameter

$$\mathcal{L}_{APA} = \frac{1}{N_p}\sum_{j_p}\mathcal{L}_{pj_p} + \frac{1}{N_w}\sum_{j_w}\mathcal{L}_{wj_w} + \frac{1}{N_u}\sum_{j_u}\mathcal{L}_{uj_u}$$

- $\mathcal{L}_{APA}$: A weighted sum of the mean square error (MSE) losses to different linguistic levels

$$\mathcal{L}_{Cor} = \ell(\hat{\Sigma}, \Sigma)$$

- MSE between the correlation matrices of predicted aspect scores ($\hat{\Sigma}$) and the corresponding target labels (Σ)

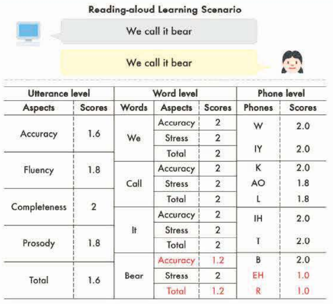
- Pre-training Strategies
  - At lower linguistic levels (i.e., phone and word levels), we leverage the mask-predict objective
  - For the utterance level, we use a strategy that predicts the relatively high or low accuracy scores for a pair of utterances

4 Experiments

- Main Results

*Acc.:Accuracy, Comp.: Completeness

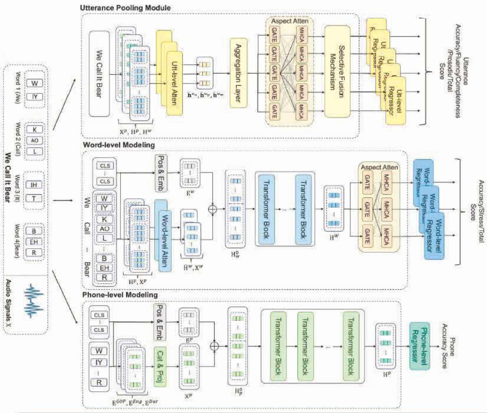
Models	Phone Score		Word Score (PCC)			Utterance Score (PCC)				
	MSE↓	PCC↑	Acc.↑	Stress↑	Total↑	Acc.↑	Comp.↑	Fluency↑	Prosody↑	Total↑
Lin2021	-	-	-	-	-	-	-	-	-	0.720
Kim2022	-	-	-	-	-	-	-	0.780	0.770	-
Ruy2023	-	-	-	-	-	0.719	-	0.775	0.773	0.743
LSTM	0.089	0.591	0.514	0.294	0.531	0.720	0.076	0.745	0.747	0.741
GOPT	0.085	0.612	0.533	0.291	0.549	0.714	0.155	0.753	0.760	0.742
HiPAMA	0.084	0.616	0.575	0.320	0.591	0.730	0.276	0.749	0.751	0.754
HierTFR	0.081	**0.644**	**0.622**	**0.325**	**0.634**	**0.735**	**0.513**	**0.801**	**0.795**	**0.764**


Reading-aloud Learning Scenario
We call it bear
We call it bear

Utterance level		Word level			Phone level	
Aspects	Scores	Words	Aspects	Scores	Phones	Scores
Accuracy	1.6	We	Accuracy	2	W	2.0
			Stress	2		
Fluency	1.8		Total	2	IY	2.0
		Call	Accuracy	2	K	2.0
			Stress	2	AO	1.8
Completeness	2		Total	2	L	1.8
		It	Accuracy	2	IH	2.0
Prosody	1.8		Stress	2		
			Total	2	T	2.0
		Bear	Accuracy	1.2	B	2.0
Total	1.6		Stress	2	EH	1.0
			Total	1.2	R	1.0

- An APA system is typically instantiated in a read-aloud scenario
- The L2 learner is presented with a text prompt and instructed to pronounce it correctly
- A running example curated from the speechocean762 dataset

3 Methodology: Hierarchical Interactive Transformer (HierTFR)



- Problem Formulation

Inputs of the proposed model:

- A time sequence of audio signals X uttered by an L2 learner
- The reference text prompt T with M words and N phones

Outputs of the proposed model:

- For each linguistic unit $g \in \{p, w, u\}$, the APA model learns to predict a set of aspect scores $\mathrm{A}^g = \{a_1^g, a_2^g, \dots, a_{N_g}^g\}$
  - p, w, u stands for the phone-, word-, and utterance-level linguistic units
  - N_g is the number of pronunciation aspects of the linguistic unit g

- Evaluation Dataset and Metrics
  - Speechocean762:

Granularity	Aspect	Score Interval	# of Counts	
			Train	Test
Phone	Accuracy	[0, 2]	47K	47K
Word	Accuracy Stress Total	[0, 10]	16K	16K
Utterance	Accuracy Completeness Fluency Prosody Total	[0, 10]	2.5K	2.5K

- Metrics:
  - **Pearson Correlation Coefficient (PCC).** Quantifying the linear correlation between predicted and ground-truth scores
  - **Mean Square Error (MSE).** Benchmarking for phone-level pronunciation accuracy in comparison with prior arts

- Ablation Studies

Models	Phone Score	Word Score			Utterance Score				
	Acc.	Acc.	Stress	Total	Acc.	Comp.	Fluency	Prosody	Total
HierTFR	**0.644**	**0.622**	**0.325**	**0.634**	**0.735**	**0.513**	**0.801**	**0.795**	**0.764**
w/o CorrLoss	0.639	0.605	0.348	0.620	0.728	0.520	0.796	0.789	0.758
w/o Pretrain	0.621	0.545	0.318	0.559	0.716	0.215	0.770	0.772	0.739
w/o SFusion	0.630	0.608	0.328	0.622	0.728	0.378	0.784	0.782	0.756
w/o AspAtt	0.636	0.584	0.290	0.596	0.724	0.383	0.784	0.775	0.746

得獎感言： 首先，感謝評審老師們對我的肯定，這份殊榮對我而言意義非凡。感謝我的指導教授—陳柏琳老師，您用以身作則的方式，為我樹立了良好研究員的典範。 此外，向我的家人和女朋友致上最深的謝意。你們是我堅實的後盾，讓我能無後顧之憂地專注於學習與研究。最後，這項肯定對我而言不僅是榮耀，更是一種提醒。我時刻地警惕自己要虛懷若谷、保持積極進取，謙虛求學。

財團法人中技社
CTCI FOUNDATION