MATRIX-STRUCTURED HIERARCHICAL CONVOLUTIONAL MODELING FOR PRONUNCIATION ASSESSMENT AND MISPRONUNCIATION DETECTION

David Fernández-García, César González-Ferreras, Valentín Cardeñoso-Payo, Mario Corrales-Astorgano

ECA-SIMM Research Group. Universidad de Valladolid (Spain)

(david.fernandez, cesargf, valentin.cardenoso, mario.corrales)@uva.es

ABSTRACT

We introduce **M3C**[1], a hierarchical, matrix-structured convolutional framework for automatic pronunciation assessment (APA) and phone-level mispronunciation detection (MDD). Heterogeneous representations–GoP (LPP/LPR), canonical-phone embeddings, self-supervised learned representations (HuBERT, Wav2Vec 2.0, WavLM), and prosodic features–are reorganized into column-aligned matrices and compressed by a self-designed CNN block (**Compact Convolutional Condenser**) across feature extraction, phone, word, and utterance levels. Multi-aspect attention is used to correlate same-level aspects, and APA is jointly trained with an auxiliary phone classifier for MDD. In a GoP-only parity setting, **M3C** outperforms *GOPT* and recent *CNN-Based* models on the *speechocean762* corpus, with the largest gains at word level (+19.4% and +7.2% respectively); with all the features, our model also improves state of the art at word level and MDD F1 (+15%). Ablation studies confirm the importance of the proposed matrix structured feature extraction and explicit triphone context, highlighting the effectiveness of proximity-driven local modeling for competitive APA.

Index Terms— computer-assisted pronunciation training, automatic pronunciation assessment, mispronunciation detection, CNN

1. INTRODUCTION

Computer-Assisted Pronunciation Training (CAPT) provides second-language (L2) learners with objective, fine-grained feedback, usually under a reading-aloud paradigm, and typically integrates two complementary components: (i) phone-level mispronunciation detection and diagnosis (MDD) and (ii) automatic pronunciation assessment (APA) that produces aspect-specific proficiency scores across multiple linguistic granularities. Recent APA systems have converged on multi-aspect, multi-granularity formulations, evaluating different aspects at the phone, word, and utterance levels within a unified architecture [1, 2].

A dominant design choice is to ground APA on Goodness-of-Pronunciation (GoP) [3] cues extracted from DNN-HMM acoustic models aligned to the prompted text. While effective for segmental (phone-level) scoring, GoP alone is insufficient to robustly capture suprasegmental phenomena (fluency, prosody, stress), motivating the inclusion of prosodic statistics (duration, energy) and self-supervised learning (SSL) representations [4]. In parallel, joint modeling of APA with MDD has gained traction because phone errors correlate with degradations in higher-level aspects; coupling the tasks can yield more consistent diagnostic and scoring signals [5, 6]. Beyond architectures and inputs, recent learning objectives explicitly preserve phone-category distinctions and score ordinality to counteract feature collapse under vanilla MSE losses [7]. These ingredients underpin the strongest approaches on the benchmark *speechocean762* corpus [8], which offers multi-rater scores at phone, word, and utterance levels and is a *de facto* testbed.

Despite these advances, most competitive systems converge in the same direction: (i) rely on attention-based encoders (which are good at modeling long-range dependencies but less effective at local dependencies [9]); (ii) treat heterogeneous phone representations by simple concatenation (blurring structured relationships among different representations); (iii) consider vowel and consonant phonemic classes as equals (combining information of really different phonemic classes at GoP-based representations [10]); (iv) leave cross-aspect interactions only at word and utterance level (which is not appropriate due to the existing relationship between mispronunciations and phone scores [8]).

To address these limitations, we present a convolutional alternative based on matrix-structured inputs. Our key idea is to *restructure* heterogeneous phone representations into matrix inputs whose columns are aligned in terms of meaning and whose rows encode complementary "views". This enables column-wise convolutions to aggregate consistent multi-view evidence for each phone without resorting to global self-attention. We also split GoP-based representations (LPP, LPR) and Canonical-phone embedding, into two feature extraction pipelines, one for vowels and the other one for consonants. Building on this feature extraction approach, we devise a hierarchical model by stacking a new convolutional module–which we called **Compact Convolutional Condenser** (**CCC**)–at the phone, word, and utterance levels, augmented then with multi-aspect attention (at all three levels) and score-restrained attention pooling (only at utterance level) in order to correlate the different aspects and predictions. Finally, we jointly train APA with an auxiliary phone classifier for MDD.

2. METHODOLOGY

We propose a multi-aspect, multi-granularity framework for APA and MDD. The overall architecture is illustrated in Figure 1. Instead of attention-based backbones, that are very common in the literature [1, 4, 6], we propose a novelty method that only uses a matrix-structured as input and convolutional neural networks (CNNs) to process them.

2.1. Dataset

We use the publicly available *speechocean762* corpus [11], a read–aloud L2 English benchmark comprising 5,000 utterances from

[1]Code at https://github.com/davidgor16/M3C.git

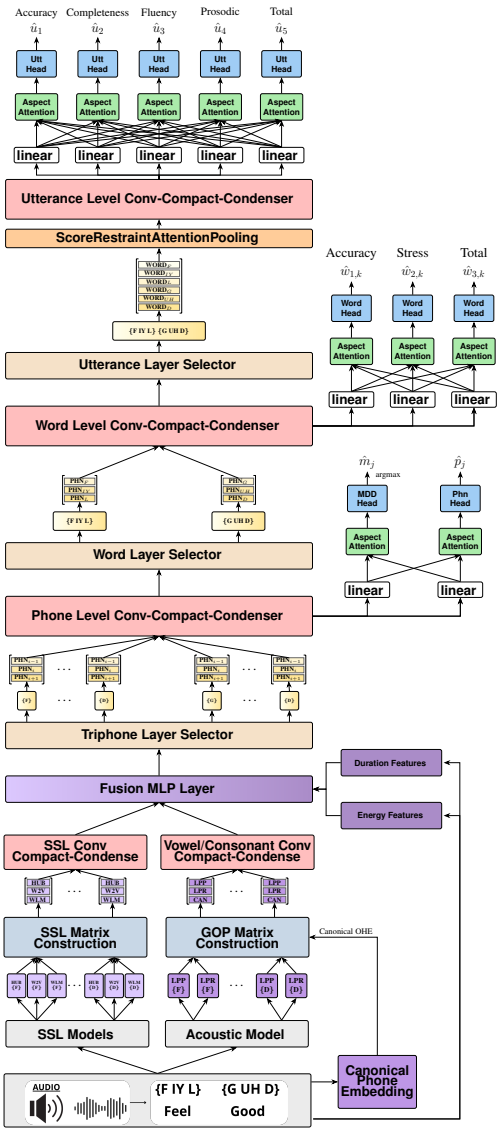

Fig. 1. Model architecture of the proposed M3C framework.

250 Mandarin L1 learners (official split: 2,500/2,500 train/test). Each item is annotated by five expert raters at three levels with multiple aspects: **utterance** (Accuracy, Completeness, Fluency, Prosody, Total; 0-10), **word** (Accuracy, Stress, Total; 0-10), and **phone** (Accuracy; 0-2). Following common practice, utterance and word level scores are linearly rescaled to the phone scale [0-2] for multi-task training. The corpus further includes phone-level transcriptions for mispronunciation diagnosis built on a set that comprises 39 phones from the CMU dictionary and 7 L2-specific phones extended with `<del>` and `<unk>` to mark deletions and unknown phone realizations.

2.2. Acoustic Input: Features Preparation

Input features are chosen following the common approach [1, 4, 6]. For each pronounced phone, we extract GoP-like scores [1] (LPP and LPR) and, as in [4], we also extract three SSL features (HuBERT [12], Wav2Vec2.0 [13] and WavLM [14]), along with simple prosodic cues: phone duration (ms) and phone energy statistics (7 values).

The input to **M3C** is built as a matrix, where each row corresponds with one view and each column refers to the same information across views. This allows us to fully exploit the relationships that these different constructions have with each other, something that would not be possible with the techniques used in other works that tend to concatenate all the representations [4, 6, 7]. In particular, two different matrices were constructed based on the specific characteristics of each set of representations: the first include LPP, LPR and the unprojected canonical-phone embbeding (CAN) –extracted using the same DNN-HMM acoustic model as [1], while the second matrix is made up of the different SSL features,

$$M_{\text{j}_{\text{gop}}} = \begin{bmatrix} \text{LPP} \\ \text{LPR} \\ \text{CAN} \end{bmatrix} \in \mathbb{R}^{3\times W}, \quad M_{\text{j}_{\text{ssl}}} = \begin{bmatrix} \text{HuBERT} \\ \text{Wav2Vec 2.0} \\ \text{WavLM} \end{bmatrix} \in \mathbb{R}^{3\times W'} \tag{1}$$

In both cases there is a positional relationship between the different representations that make up the two matrices. In the case of the M_{gop}, the different positions refer to the different kind of information that relates the pronounced phone with each of the other possible phones. In the case of the M_{ssl}, although the features learned by each model are not identical, shared pre-training objectives tend to align representations at the layer/subspace level, yielding a stable positional correspondence across the vectors [15, 16].

Before being fed into the model, M_{gop} is transformed into a vowel or consonant specific matrix, using a fixed mapping. If the pronounced phone is a vowel, then M_{gop} will only contain information about the relationship between the pronounced vowel phone and the different existing vowel phones (15 vowels in phone set). Same process is applied if the phone is a consonant (24 consonants in phone set).

$$M_{\text{j}_{\text{gop}}} = \begin{cases} M_{\text{j}_{\text{gop}}}[:, :, \mathcal{V}_{\text{vow}}] & \text{if } j \text{ is a vowel,} \\ M_{\text{j}_{\text{gop}}}[:, :, \mathcal{V}_{\text{con}}] & \text{if } j \text{ is a consonant.} \end{cases} \tag{2}$$

where, $M_{\text{j}_{\text{gop}}}$ denotes the original matrix representation, while $\mathcal{V}_{\text{vow}}$ and $\mathcal{V}_{\text{con}}$ correspond to the index sets of vowel and consonant phones, respectively.

Lastly, to prevent this information from being lost throughout the process, in the fusion phase where we merge the different final representations, a position is added to the embedding to indicate whether the representation refers to a vowel or a consonant.

2.3. Compact Convolutional Condenser (CCC)

We use a compact convolutional compressor to translate short context windows into compact representations throughout the different hierarchy levels (phone, word, and utterance), and also at feature extraction level. The objective of this block is to compress and condense information column-wise, given the relationship described in Section 2.2.

Formally, for a single matrix input representation $\mathbf{M} \in \mathbb{R}^{1\times H\times W}$, having only one input channel, and being H the number of different representations and W the representations width, we compute:

$$\mathbf{m} = \text{Conv2D}_{H\times 1}(\mathbf{M}) \tag{3}$$

$$\mathbf{p} = \text{Dropout}_{\text{cnn}}(\text{ReLU}(\text{LayerNorm}(\text{Flatten}(\mathbf{m})))) \tag{4}$$

$$\mathbf{h} = \text{Dropout}_{\text{mlp}}(\text{ReLU}(\text{LayerNorm}(\mathbf{W}\mathbf{p} + \mathbf{b}))) \tag{5}$$

Category	Model	Phone Score		Word Score (PCC)			Utterance Score (PCC)					MDD		
		MSE↓	PCC↑	Acc.↑	Stress↑	Total↑	Acc.↑	Comp.↑	Fluency↑	Prosody↑	Total↑	Prec.↑	Rec.↑	F1↑
Baseline	GOPT [1]	0.085 ± 0.001	0.612 ± 0.003	0.533 ± 0.004	0.291 ± 0.030	0.549 ± 0.002	0.714 ± 0.004	0.155 ± 0.039	0.753 ± 0.008	0.760 ± 0.006	0.742 ± 0.005	-	-	-
CNN-Based	HiPAMA [2]	0.084 ± 0.001	0.616 ± 0.004	0.575 ± 0.004	0.320 ± 0.021	0.591 ± 0.004	0.730 ± 0.002	0.276 ± 0.177	0.749 ± 0.001	0.751 ± 0.002	0.754 ± 0.002	-	-	-
	Gradformer [17]	0.079 ± 0.001	0.646 ± 0.004	0.598 ± 0.006	0.334 ± 0.013	0.614 ± 0.006	0.732 ± 0.005	0.318 ± 0.139	0.769 ± 0.006	0.767 ± 0.004	0.756 ± 0.003	-	-	-
	Bfhaformer [18]	0.080 ± 0.0001	0.646 ± 0.003	0.621 ± 0.005	**0.386** ± 0.024	0.635 ± 0.003	0.737 ± 0.004	**0.488** ± 0.136	0.770 ± 0.004	0.770 ± 0.004	0.760 ± 0.003	-	-	-
	Attention-CNN [19]	0.081 ± 0.001	0.639 ± 0.004	0.585 ± 0.005	0.269 ± 0.031	0.600 ± 0.005	0.727 ± 0.003	0.277 ± 0.082	0.766 ± 0.002	0.762 ± 0.003	0.754 ± 0.008	-	-	-
	M3C	**0.074** ± 0.001	**0.676** ± 0.008	**0.666** ± 0.014	0.297 ± 0.033	**0.676** ± 0.015	**0.753** ± 0.011	0.329 ± 0.031	**0.781** ± 0.008	**0.782** ± 0.008	**0.779** ± 0.008	79.7% ± 0.001%	78.9% ± 0.01%	79.2% ± 0.001%
SOTA	HMAMBA [6]	**0.062** ± 0.000	**0.739** ± 0.000	0.708 ± 0.000	**0.366** ± 0.000	0.718 ± 0.000	**0.807** ± 0.000	**0.278** ± 0.000	**0.848** ± 0.000	**0.843** ± 0.000	**0.829** ± 0.000	64.3% ± 0.0%	63.4% ± 0.0%	63.8% ± 0.0%
	M3C	0.066 ± 0.001	0.716 ± 0.004	**0.710** ± 0.004	0.340 ± 0.028	**0.721** ± 0.005	0.791 ± 0.003	0.268 ± 0.065	0.830 ± 0.001	0.832 ± 0.001	0.816 ± 0.002	**79.7%** ± 0.008%	**78.2%** ± 0.04%	**78.8%** ± 0.005%

Table 1. Performance comparison of M3C with baseline, CNN-based, and state-of-the-art (SOTA) systems for APA and MDD. For fair comparison, baseline and CNN-based models use only GoP features, while SOTA models additionally incorporate SSL and prosodic cues. Best results within each feature group are highlighted in bold. Standard deviations are unavailable for HMAMBA [6], so all values are 0.

where $\mathbf{W} \in \mathbb{R}^{(C \cdot W) \times d}$, C is the number of convolutional filters, and d is the target dimensionality. This design captures local interactions while preserving positional relations, as position n now contains a condensed value for a concrete information.

2.4. Hierarchical Modeling Approach

2.4.1. Feature Extraction Level

Since we have three different representation levels (M_{gop}, M_{ssl} and Prosodic features), first thing is to extract the most relevant information of each representation to later fuse them all in a single tier. To do that, three **CCCs** are used. First two will extract M_{gop} features, one to be used if the uttered phone is a vowel and the other if it is a consonant. The other one will process M_{ssl}.

$$\mathbf{h_{gop}}^{(\mathbf{j})} = \begin{cases} \mathrm{CCC^{vowels}}_{3\times1}(\mathbf{M_{gop}}^{(\mathbf{j})}) & \text{if } j \text{ is a vowel,} \\ \mathrm{CCC^{consonants}}_{3\times1}(\mathbf{M_{gop}}^{(\mathbf{j})}) & \text{if } j \text{ is a consonant.} \end{cases} \tag{6}$$

$$\mathbf{h_{ssl}}^{(\mathbf{j})} = \mathrm{CCC^{ssl}}_{3\times1}(\mathbf{M_{ssl}}^{(\mathbf{j})}) \tag{7}$$

After both feature extraction processes, we concatenate the output of the **CCCs** with the Prosodic Features. As a final step, this new representation is projected using and MLP Fusion Layer in order to integrate the information of the three sources. As result, we obtained a fused phone-level representation (x_{j})

2.4.2. Phone-Level

A local context is built around each phone by stacking the previous, current, and next phones (triphone window), applying padding to the matrix when needed. Then, **CCC** with kernel size 3x1 operates across these rows while preserving the positional semantics described above.

$$\mathbf{M}_{\text{phn}}^{(j)} = \mathrm{Pad}_3\left(\begin{bmatrix} \mathbf{x}_{j-1} \\ \mathbf{x}_j \\ \mathbf{x}_{j+1} \end{bmatrix}\right) \tag{8}$$

$$\mathbf{h_{phn}}^{(\mathbf{j})} = \mathrm{CCC^{phn}}_{3\times1}(\mathbf{M_{phn}}^{(\mathbf{j})}) \tag{9}$$

After that, an aspect-attention module is applied [2], which allows to capture the existing correlation between the phone-score and MDD aspects before prediction, since a relationship between this two aspects exists [11].

$$\mathbf{h}_r^{(j)} = \mathbf{W}_r\, \mathbf{h}^{(j)} + \mathbf{b}_r \qquad r \in \{score, mdd\} \tag{10}$$

$$\mathbf{a}_r^{(j)} = \mathrm{AspectAttn}\big(\mathbf{h}_r^{(j)},\, \mathbf{h}_{r'}^{(j)}\big) \qquad r' \neq r. \tag{11}$$

Finally, two MLP heads are applied on top: (1) a regressor that outputs the phone accuracy score, and (2) a 48-way classifier for MDD.

2.4.3. Word-Level

Triphone representations belonging to the same word are stacked and padded (if necessary) to 12 phones (rows) matrices. We stack the different phones in 12-row matrices because 12 is the maximum number of phones that a word has within the *speechocean762* corpus [11]. Then, following the same structure as the one seen at phone-level, **CCC** with kernel size 12x1 is applied to obtain compact word representation. After that, three aspect-attention blocks followed by three MLP layers are used to predict word aspects.

2.4.4. Utterance-Level

We stack word-level representation (per phone) across the utterance and apply a **CCC** with kernel size 50x1 (50 is the maximum number of phones that a sentence can have in *speechocean762* Corpus). In this case, we adopt score-restrained attention pooling [20], such that the predicted phone and word scores provide salience weights that reweight utterance hidden states before the final regressors.

$$\mathbf{p}_{utt} = \mathrm{SRAPool}\Big(\mathbf{h}_{utt},\, \hat{\mathbf{p}},\, \{\hat{\mathbf{w}}_r\}_{r=1}^{3}\Big), \tag{12}$$

Then, the same approach as in the lower levels is followed to predict utterance-level aspects.

2.5. Optimization

Our model is trained in a multi-task learning setup, combining in a weighted manner the loss functions of APA and MDD.

$$\mathcal{L} = \mathcal{L}_{\text{APA}} + \beta\, \mathcal{L}_{\text{MDD}} \qquad (\beta = 0.03) \tag{13}$$

Models	Phone-level Score		Word-level Score (PCC)			Utterance-level Score (PCC)					MDD		
	MSE↓	PCC↑	Acc.↑	Stress↑	Total↑	Acc.↑	Comp.↑	Fluency↑	Prosody↑	Total↑	Prec.↑	Rec.↑	F1↑
M3C	**0.066** ± 0.001	**0.716** ± 0.004	0.710 ± 0.004	**0.340** ± 0.028	0.721 ± 0.005	**0.791** ± 0.003	0.268 ± 0.065	0.830 ± 0.001	0.832 ± 0.001	**0.816** ± 0.002	79.7% ± 0.008%	78.2% ± 0.04%	78.8% ± 0.005%
w/o Matrix Feature Extraction	0.082 ± 0.001	0.631 ± 0.008	0.627 ± 0.004	0.240 ± 0.020	0.638 ± 0.004	0.779 ± 0.002	0.085 ± 0.013	0.814 ± 0.003	0.814 ± 0.003	0.798 ± 0.001	76.6% ± 0.008%	75.5% ± 0.012%	75.8% ± 0.007%
w/o Triphones	0.121 ± 0.015	0.611 ± 0.071	0.635 ± 0.073	0.282 ± 0.054	0.647 ± 0.071	0.681 ± 0.152	0.178 ± 0.095	0.694 ± 0.197	0.692 ± 0.196	0.696 ± 0.169	57.4% ± 0.278%	56.2% ± 0.287%	54.9% ± 0.306%
w/o Phone Aspect Attention	0.069 ± 0.001	0.702 ± 0.005	**0.711** ± 0.007	0.320 ± 0.027	**0.722** ± 0.006	0.790 ± 0.004	**0.300** ± 0.061	**0.835** ± 0.003	**0.836** ± 0.003	0.815 ± 0.003	**80.3%** ± 0.000%	**79.2%** ± 0.002%	**79.6%** ± 0.002%

Table 2. Ablation study of the proposed method on APA and MDD.

For the APA branch, we minimize Mean Squared Error (MSE) at three granularities (phone, word and utterance). The APA losses are computed by directly summing the averaged losses per granularity:

$$\mathcal{L}_{\text{APA}} = \mathcal{L}_{\text{phone}} + \mathcal{L}_{\text{word}} + \mathcal{L}_{\text{utterance}}. \quad (14)$$

where $\mathcal{L}_{\text{utterance}}$ and $\mathcal{L}_{\text{word}}$ are averaged utterance and word level losses of five utterance-level labels and three word-level labels, respectively; $\mathcal{L}_{\text{phone}}$ is the phone loss.

For the MDD branch, we optimize the cross-entropy over total number of phones (N_{phn}):

$$\mathcal{L}_{\text{MDD}} = \frac{1}{N_{\text{phn}}} \sum_{j=1}^{N_{\text{phn}}} \text{CE}(\hat{\mathbf{z}}_j,\, y_j)\,, \quad y_j \in \{1, \ldots, 48\}. \quad (15)$$

3. EXPERIMENTS

3.1. Implementation Details

The proposed model uses a global hidden size of 30 and 32 convolutional filters per CCC. The model was trained with Adam optimizer (learning rate 1×10^{-3}, batch size 2) for 50 epochs, using dropout regularization. Evaluation was based on Pearson correlation coefficient (PCC) and mean squared error (MSE), averaged over five runs with different random seeds.

3.2. Results

In Table 1, we compare our method with three model families: *baseline*, *CNN-based*, and *state-of-the-art*. For parity with *baseline* and *CNN-based* systems, which use only GoP inputs, we ablated SSL and prosodic features from our model. Compared with the baseline GOPT [1], our approach improves performance at all granularities, with a +19.4% relative gain at the word level and a +12.3% overall increase across aspects. Compared with recent *CNN-based* architectures (often paired with attention), our method—*to our knowledge, the only one that uses convolutional layers exclusively as core processing modules*—achieves consistent gains: +7.2% (word), +6.1% (phone), and +2.2% (utterance) mean relative improvements. Because prior *baseline* and *CNN-based* works did not include MDD, we omit that aspect from this comparison. Relative to the current state of the art, HMAMBA [6], which integrates a single convolutional block at the word level, our approach yields an advantage at exactly that level and further improves MDD with a +15% F1 increase.

3.3. Ablation Studies

Table 2 summarizes the results of ablating three principal characteristics of **M3C**: (i) matrix-structured feature extraction (we concatenate all features, like mainly all works do), (ii) explicit triphone context, and (iii) phone-level aspect attention. Removing matrix feature extraction consistently hurts performance (e.g., phone PCC 0.716→0.631; word-total PCC 0.721→0.638; MDD F1 78.8→75.8), reinforcing the contribution of our feature extraction pipeline. Eliminating triphone contexts yields the biggest and more variable drop (phone MSE 0.066→0.121; word-level total PCC 0.721→0.647; utterance-level Total PCC 0.816→0.696; MDD F1 78.8→54.9), underscoring the importance of a close local phoneme context; the standard deviation increase by an order of magnitude (phone PCC ±0.004→±0.071; word-level total PCC ±0.005→±0.071; utterance-level Total ±0.002→±0.169; MDD F1 ±0.005→±0.306) revealing markedly unstable training and poorer generalization without triphone context. Finally, ablating phone-aspect attention yields minor changes: some drop at phone-level (phone PCC 0.716→0.702), almost no changes at word and sentence level, and a slight increase in performance for MDD which is not enough to compensate the phone-level performance decay. Overall, the performance gains are mainly attributable to matrix structuring and explicit triphone modeling, which also propagate benefits to word and utterance level scoring.

4. CONCLUSIONS

Our results demonstrate that organizing inputs in a matrix structure and employing the **CCC** as the principal processing block consistently outperforms strong convolutional-based models, validating both the new block and the overall modeling strategy. Our architecture enables more faithful modeling of local relationships; combined with our empirical results, this highlights the importance of locality in APA–often downplayed in attention-centric work. Local context dominance is intuitive: pronunciation influences are primarily proximity-driven (e.g., triphones or within-word phone interactions). With a local yet incremental design, the model also accumulates global dependencies, explaining the strong utterance-level results.

Compared with the state of the art [6], the sizable MDD improvement stems from our feature-extraction pipeline and the explicit triphone representation (as can be seen in Table 2): the offset of the preceding phone and the onset of the following one jointly constrain the central phone, and their aggregation enhances detection. Finally, gains in *Word Accuracy* and *Word Total* arise from constructing condensed, word-level triphone embeddings and jointly aggregating the triphones within each word. This design naturally encodes cross-word boundary cues (end of the previous word and start of the next), which are highly informative at the word granularity.

5. ACKNOWLEDGEMENTS

This work was carried out in Project PID2021-126315OB-I00 that was supported by MCIN / AEI / 10.13039/501100011033 / FEDER, EU.

6. REFERENCES

[1] Yuan Gong, Ziyi Chen, Iek-Heng Chu, Peng Chang, and James Glass, "Transformer-based multi-aspect multi-granularity non-native English speaker pronunciation assessment," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7262–7266.

[2] Heejin Do, Yunsu Kim, and Gary Geunbae Lee, "Hierarchical pronunciation assessment with multi-aspect attention," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[3] Silke M Witt and Steve J Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[4] Fu-An Chao, Tien-Hong Lo, Tzu-I Wu, Yao-Ting Sung, and Berlin Chen, "3M: An effective multi-view, multi-granularity, and multi-aspect modeling approach to English pronunciation assessment," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 575–582.

[5] Yue-Yang He, Bi-Cheng Yan, Tien-Hong Lo, Meng-Shin Lin, Yung-Chang Hsu, and Berlin Chen, "JAM: A unified neural architecture for joint multi-granularity pronunciation assessment and phone-level mispronunciation detection and diagnosis towards a comprehensive capt system," in *Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2024, pp. 1–6.

[6] Fu-An Chao and Berlin Chen, "Towards efficient and multifaceted computer-assisted pronunciation training leveraging hierarchical selective state space model and decoupled cross-entropy loss," in *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, Apr. 2025, pp. 1947–1961.

[7] Bi-Cheng Yan, Yi-Cheng Wang, Jiun-Ting Li, Meng-Shin Lin, Hsin-Wei Wang, Wei-Cheng Chao, and Berlin Chen, "ConPCO: Preserving phoneme characteristics for automatic pronunciation assessment leveraging contrastive ordinal regularization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[8] Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang, "speechocean762: An open-source non-native English speech corpus for pronunciation assessment," *arXiv preprint arXiv:2104.01378*, 2021.

[9] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech 2020*, 2020, pp. 5036–5040.

[10] Mostafa Shahin, Beena Ahmed, Jim X. Ji, and Kirrie Ballard, "Anomaly detection approach for pronunciation verification of disordered speech using speech attribute features," in *Interspeech 2018*, 2018, pp. 1671–1675.

[11] Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang, "speechocean762: An open-source non-native English speech corpus for pronunciation assessment," in *Interspeech*, 2021, pp. 3710–3714.

[12] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, pp. 3451–3460, Oct. 2021.

[13] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 12449–12460.

[14] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.

[15] Anton de la Fuente and Dan Jurafsky, "A layer-wise analysis of Mandarin and English suprasegmentals in SSL speech models," in *Interspeech 2024*, 2024, pp. 1290–1294.

[16] Ankita Pasad, Bowen Shi, and Karen Livescu, "Comparative layer-wise analysis of self-supervised speech models," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[17] Hao-Chen Pei, Hao Fang, Xin Luo, and Xin-Shun Xu, "Gradformer: A framework for multi-aspect multi-granularity pronunciation assessment," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 554–563, 2024.

[18] Wenxu Du, Aishan Wumaier, Yahui Shi, Nian Yi, and Dehua Liu, "A multi-aspect multi-granularity pronunciation assessment method based on branchformer encoder and hierarchical aggregation," in *MultiMedia Modeling*, 2025, pp. 16–29.

[19] Jianlei Yang, Aishan Wumaier, Zaokere Kadeer, Liejun Wang, Shen Guo, and Jing Li, "Attention-CNN combined with multi-layer feature fusion for English L2 multi-granularity pronunciation assessment," in *IEEE 4th International Conference on Pattern Recognition and Machine Learning (PRML)*, 2023, pp. 449–457.

[20] Fu-An Chao, Tien-Hong Lo, Tzu-I Wu, Yao-Ting Sung, and Berlin Chen, "A hierarchical context-aware modeling approach for multi-aspect and multi-granular pronunciation assessment," in *Interspeech*, 2023, pp. 974–978.