ENHANCING AUTOMATIC SPEECH ASSESSMENT LEVERAGING HETEROGENEOUS FEATURES AND SOFT LABELS FOR ORDINAL CLASSIFICATION

Wen-Hsuan Peng[1], *Sally Chen*[2], *Berlin Chen*[1]

[1]Department of Computer Science and Information Engineering, National Taiwan Normal University
[2]The Language Training & Testing Center
{61147006s, berlin}@ntnu.edu.tw
sallychen@lttc.ntu.edu.tw

2024 IEEE Spoken Language Technology Workshop (SLT) | 979-8-3503-9225-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/SLT61566.2024.10832269

ABSTRACT

The general goal of automated speech assessment (ASA) is to provide a consistent and objective evaluation on the spoken language proficiency of an L2 learner or test-taker. In contrast to most previous work that treats ASA as a nominal multi-classification task and thus neglects the sequential nature of proficiency grades, this paper explores the notion of soft labels for use in ASA. In particular, we strive to enhance ASA performance by examining two critical issues: (1) the impact of applying soft labels instead of hard labels in the optimization of ordinal classification for ASA, and (2) the effects of combining self-supervised learning (SSL) with handcrafted indicator features via a novel modeling paradigm. Our results demonstrate that the proposed model can considerably enhance performance compared to existing strong baselines. The improvement is evident not only in the test dataset of seen prompts but also in those of unseen prompts, suggesting the robust generalization and adaptability of our method.

Index Terms— Automated speech assessment, Multimodal model, End-to-end neural network

1. INTRODUCTION

Automated speech assessment (ASA) evaluates and quantifies the proficiency of foreign language learners or test-takers by assigning scores on their spoken responses. By offering immediate and objective feedback, ASA systems are anticipated to better ensure consistent and unbiased scoring so as to assist experienced human raters. Recent breakthroughs in spoken language technologies have spurred a host of research on ASA. From neural architectures to current self-supervised learning techniques [1] [2] which leverage pre-trained models, these research endeavors have greatly enhanced ASA performance. Furthermore, prior studies incorporating aspect-related features [3] into an ASA system have been shown to further increase the overall effectiveness. Nevertheless, most current ASA models still suffer from certain limitations. Notably, while ASA scores intrinsically have a sequential nature, many prior arts simply approach ASA as a nominal multi-classification task, using hard labels instead of acknowledging the sequential order.

On these grounds, this paper aims to improve ASA performance by (1) investigating the benefits of using soft labels to the optimization of ordinal classification, and (2) exploring the integration of self-supervised learning (SSL) with handcrafted indicator via a novel modeling paradigm, in order to examine their combined effect on model efficacy.

Our results show that optimizing ordinal classification through the synergy effect of soft-label optimization and self-supervised learning (SSL) combinations with handcrafted features can aptly enhance performance compared to existing strong baselines. The improvement is evident not only in the test set consisting of seen question prompts but also in the test set consisting of unseen question prompts, indicating the robust generalization and adaptability of our approach.

2. RELATED WORK

Early studies in ASA mainly focused on pronunciation quality [4], comparing the learner's digital representation of a scripted response to a model of the expected native speaker phoneme pronunciation distribution. As the field evolved, researchers began to broaden their scope to include more comprehensive aspects of language proficiency, investigating methods for automatically assessing spontaneous speech, also known as automated speech assessment (ASA) or automated speech proficiency assessment (ASPA) [5]. Traditional ASA models typically involve automatic speech recognition (ASR) so as to generate time-aligned word sequences for an input speech. This process is broadly divided into two parts: fluency features extraction and scoring models. The commonly-used fluency includes long silence, words per second, phone, and others [6]. Leveraging machine learning techniques, ASA has integrated statistical scoring models, including the support vector machine (SVM), Gaussian and linear regression models and their variants to predict fluency scores. While these well-established models often surpass traditional goodness of pronunciation (GOP) score-based

models in performance, feature engineering and selection might be time-consuming and labor-intensive. Deep learning methods have shown impressive performance on various spoken language processing tasks such as ASR, emotion recognition, keyword spotting, and speaker identification, and these advancements have recently been extended to ASA [7][8][9][10][11]. In these studies, pre-trained models have been employed to construct contextual representations for spoken responses. Specifically, it has been demonstrated that such models can capture speech-related features and linguistic information, such as acoustics, fluency, pronunciation, and text-based syntactic and semantic features for L1 and L2 learners [12]. [1] reported the usefulness of Wav2vec2.0 in building an ASA system. Further research has explored the integration of BERT and Wav2Vec2.0 in proficiency assessment [13]. Park et al. [3] also demonstrated the capability of Transformer-based architectures in predicting proficiency across multiple sub-levels (delivery, language use, and topic development) as well as providing a holistic score, leveraging both speech and text representations.

Building on the aforementioned observations, we in this paper design and implement a novel ASA model which can better render three salient aspects, viz. content, delivery, and languages, as well as their intricate interactions for ASA, dubbed SAMAD (Speech Assessment with Multi-Aspect Design). Notably, we also explore the synergy effect of soft-label optimization and self-supervised learning (SSL) in conjunction with handcrafted features for further performance improvements.

3. DATASET

In this study, we employed in-house datasets compiled and curated from the General English Proficiency Test (GEPT) for our experiments. The General English Proficiency Test, developed and administered by the Language and Testing Center (LTTC), targets English learners at all levels in Taiwan, covering four language skills of listening, reading, writing, and speaking.

There are several reasons why we considered using the in-house dataset. First, to the best of our knowledge, datasets featuring non-native spontaneous English speech are scarce. Most existing ASA research relies on proprietary datasets maintained by prominent testing organizations, which are typically inaccessible to external researchers. Secondly, among the limited public resources, the ICNALE dataset is constrained by its reliance on scores derived from standardized tests (e.g., TOEFL, TOEIC, IELTS, and others) and lacks evaluations based on responses to specific prompts, limiting its applicability for more detailed analysis. Lastly, the GEPT dataset, commonly encountered in real educational scenarios, features a single holistic score and exhibits a normal distribution. Although this makes the tasks more challenging, it offers a valuable opportunity to test whether our model can be implemented in realistic situations. Considering the above reasons, we decided to use the GEPT dataset in our research. We utilized four datasets from the Intermediate level of the GEPT speaking tests, each dataset consisting of three test scenarios: reading aloud, question answering, and picture description. Each response was rated by at least two experts. A third score is assigned if the scores given by these two experts differ by one grade level. The final adjudicated scores are used as the reference scores to build the scoring model. Our research specifically targets the picture description task due to its substantial duration and rich content. In the picture description task, the test takers observe a picture related to a non-academic topic (e.g., Where was this picture probably taken?) and are expected to provide a detailed response based on the prompt accompanying the image. The responses in the Intermediate dataset are approximately 85 seconds long and contain about 121 words on average.

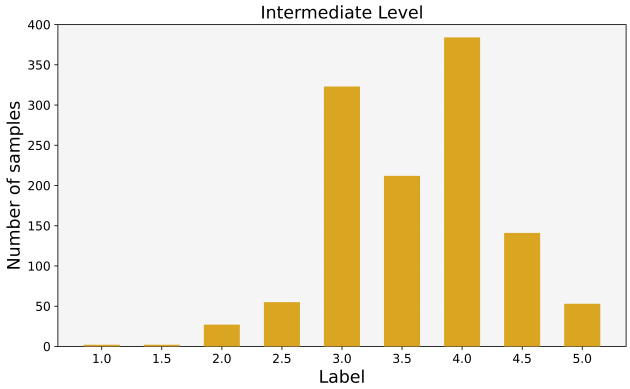

Fig. 1. Class distribution of the Intermediate dataset.

The annotations of the GEPT dataset utilize a distinct scoring policy, rating each response on a scale from 0 to 5, which differs from the CEFR scoring scale, as shown in Figure 1. However, LTTC's studies have demonstrated that the GEPT evaluation process aligns with internationally accepted CEFR standards [14]. Achieving a score above 3 at the intermediate level of the GEPT indicates a proficiency above CEFR B1 level, while scores below or equal to 3 indicating failing the test. In the studies, ASA tasks are treated as classification tasks. However, the GEPT rating scale produces continuous values by averaging each human rater's scores. Therefore, to establish discrete classes suitable for the classification task, we converted the ratings into integers using a rounding-down method. Specifically, labels such as 3.5 were treated as 3 during both the training and testing phases. This adjustment aligns with the real testing scenario where a test taker receiving a score of 3.5 is considered to have failed the GEPT Intermediate level test. As illustrated in Table 1, the GEPT dataset was divided into four subsets: training (N=719), development (N=90), test (N=90), and unseen test (N=300), facilitating comprehensive evaluation on both seen and unseen prompts. This division not only as-

sesses the model's generality across speaker traits but also its adaptability to new scenarios.

Scale	Train	Dev.	Seen Test	Unseen Test	Total
1	4	0	0	0	4
2	61	6	11	4	82
3	317	39	37	142	535
4	310	39	39	137	525
5	27	6	3	17	53
Total	719	90	90	300	1199

Table 1. Number of spoken responses in the GEPT datasets.

4. METHODOLOGY

The proposed SAMAD model can render three salient aspects, viz. content, delivery, and languages, as well as their intricate interactions for ASA. In particular, SAMAD is also equipped with two innovative elements: optimization with soft labels to emphasize the ordinal properties of ASA tasks, and the combination of self-supervised pre-trained (SSL) with handcrafted indicator features to enrich the representation of a spoken response.

4.1. Feature Extractions

Prompt-Relevance Module. In the prompt-relevance module, we utilize a pre-trained BERT model as the feature extractor. Specifically, an input spoken response is captured and transcribed by an ASR system and in turn paired with the prompt as input to this module.

Delivery Module. The delivery features encompass acoustic and fluency features, which are derived from established methodologies suggested by [15] and [12]. These features include pitch, intensity, pauses, silences, duration, and among others.

Language Use Module. We initially convert an input spoken response into its corresponding word sequence via an ASR system and concatenate multiple one-hot vectors representing part-of-speech (POS), syntactic dependency labels (DEP), and morphology (Morphs.) features from spaCy[1]. In total, there are 21, 45, and 181 word-level labels for POS, DEP, and Morphs.

4.2. Overall Architecture

Three major aspects in our proposed ASA model are prompt-relevant (C), delivery (D) and language use (L). We denoted with $X_{\{C,D,L\}} \in \mathbb{R}^{T_{\{C,D,L\}} \times d_{\{C,D,L\}}}$ the input features sequences from these three aspects. With these notations, this subsection will detail the components of our SAMAD model, whose processing flow is divided into three parts stacked in tandem: temporal convolution, cross-aspect attention, and the output layer.

Temporal Convolutions. To provide the input sequence with a temporal dimension, we pass the input sequences through a 1D temporal convolution layer [16]:

$$\hat{X}_{\{C,D,L\}} = \text{Conv1D}(X_{\{C,D,L\}}, k_{\{C,D,L\}}) \in \mathbb{R}^{T_{\{C,D,L\}} \times d} \quad (1)$$

where $k_{\{C,D,L\}}$ are the sizes of the convolutional kernels for aspects $\{C, D, L\}$, and d is a common dimension. The convolved sequence are expected to contain the local structure of the sequence. Furthermore, since the temporal convolution projects the features of different aspects to the same dimension d, the dot-product operation is performed in the cross-aspect attention module.

Multi-Head Self-Attention Mechanism. To effectively examine each point of the long-range response, we employ a multi-head self-attention mechanism for the response-level representation. The multi-head self-attention mechanism is formulated as follows. First, for each head i and each aspect $A \in \{C, D, L\}$, the input sequence is projected into the query matrix $Q_{i,A}$, keys matrix $K_{i,A}$, and values matrix $V_{i,A}$:

$$Q_{i,A} = X_A W^Q_{i,A} \quad (2)$$

$$K_{i,A} = X_A W^K_{i,A} \quad (3)$$

$$V_{i,A} = X_A W^V_{i,A} \quad (4)$$

where X_A represents the input matrix for aspect A, and $W^Q_{i,A}$, $W^K_{i,A}$, and $W^V_{i,A}$ are the parameter matrices for the query, key, and value matrices, respectively, associated with aspect A.

Next, compute the attention scores for each head i and aspect A using:

$$\text{Attention}(Q_{i,A}, K_{i,A}, V_{i,A}) = \text{softmax}\left(\frac{Q_{i,A}K^\top_{i,A}}{\sqrt{d_{k,A}}}\right) V_{i,A} \quad (5)$$

where $d_{k,A}$ is the dimensionality of the keys for aspect A, used to scale the dot products, ensuring stable gradients during training.

Finally, concatenate the outputs from all heads for each aspect A:

$$Z_A = [head_{1,A}; \ldots; head_{h,A}]W^O_A \quad (6)$$

where $head_{i,A} = \text{Attention}(Q_{i,A}, K_{i,A}, V_{i,A})$ and W^O_A is another parameter matrix specific to aspect A.

Cross-aspect Attention Module. The next step is the cross-aspect attention module, which allows the module to focus dynamically on features across aspects. In the following, we pass delivery (D) to content (C), marked as "$D \rightarrow C$," which can be expressed by

$$Y_C = \text{CM}_{D\rightarrow C}(Z_D, Z_C) \quad (7)$$

$$= \text{softmax}\left(\frac{Q_C K^\top_D}{\sqrt{d_k}}\right) V_D \quad (8)$$

$$= \text{softmax}\left(\frac{Z_C W_{Q_C} W^\top_{K_D} Z^\top_D}{\sqrt{d_k}}\right) Z_D W_{V_D} \quad (9)$$

[1] https://spacy.io/

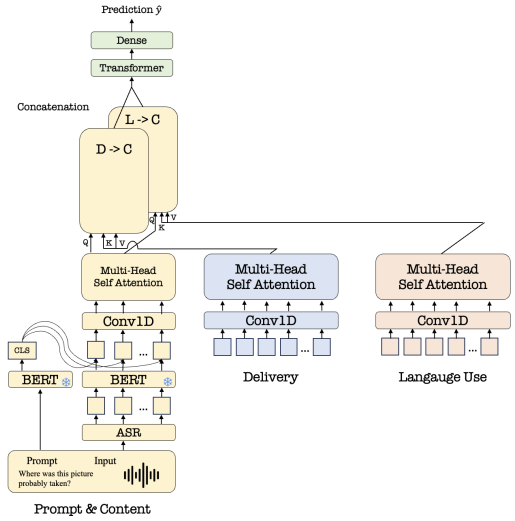

Fig. 2. A schematic depiction of the Speech Assessment with Multi-Aspect Design (SAMAD) model that incorporates three types of input.

After the cross-attention module, a residual connection and layer normalization are implemented to incorporate the original modality from the other modality:

$$Z_{D\to C} = LN(Y_C + Z_C) \tag{10}$$

After applying layer normalization (LN), a feed-forward layer is utilized to combine the feature representations:

$$Z_{D\to C} = LN(Z_{D\to C} + FFN(Z_{D\to C})) \tag{11}$$

Likewise, we can get the representation vector $Z_{L\to D}$ from language use (L) to content (C).

Self-Attention Transformers and Prediction. Finally, we concatenate the outputs from the cross-aspect Transformers that share the same target aspect, producing $Z_{\{C,D,L\}} \in \mathbb{R}^{T_{\{C,D,L\}}\times 2d}$, $Z_C = [Z_{\{D\to C\}}; Z_{\{L\to C\}}]$. Z_C, is then passed through a self-attention Transformer to gather temporal information for making predictions. The last elements of the sequence models are extracted and passed through fully-connected layers to make the final predictions.

4.3. Soft Labels for Ordinal Classification

ASA can be framed as an ordinal classification task, with labels spanning from beginner to advanced learners (e.g., Grade 0 to Grade 5). Previous studies often overlook the fact that a misclassification of a Grade 2 response as Grade 4 one should be penalized more severely than a misclassification as Grade 3 one. As such, we in this paper acknowledge this important trait of ASA by adopting soft-label optimization for training SAMAD, which not only enables fine-grained classification but also ensure the misclassification penalty reflects the distance between the predicted and reference grades. Although similar attempts have been made in readability assessment [17] [18], however to the best of our knowledge, there is a dearth of work investigating its effectiveness of the non-occurring words for ASA.

A bit more terminology: the definition of the *soft label* in the context of an ordinal classification task with K categories is as follows

$$y_i = \frac{\exp(-\phi(r_i, r_t))}{\sum_{k=1}^{K}\exp(-\phi(r_k, r_t))}, \tag{12}$$

where $r_i \in \mathcal{Y} = \{r_1, r_2, ..., r_K\}$ is the i-th category, r_t is the true category and $\phi(r_i, r_t)$ is a distance metric between two categories. The distance metric $\phi(r_i, r_t)$ is defined by the following piece-wise constant function:

$$\phi(r_i, r_t) = \begin{cases} 0, & i = t \\ c, & |i - t| = 1, \\ +\infty, & \text{otherwise} \end{cases} \tag{13}$$

where the positive hyper-parameter c is the distance between the actual label and its neighboring labels. To better optimize the ASA performance, we empirically determined the optimum value of c to be 1.2, as illustrated in Figure 4.

5. EXPERIMENT AND RESULTS

5.1. Baseline Model

To evaluate the performance level of the SAMAD model, we selected two iconic Transformer-based models for comparison, BERT and Wav2vec2.0, respectively, as a benchmark in text and audio processing. Additionally, we replicated the Qian2019 model [15], which incorporates various features, and the multimodel model [13], which explores the integration of BERT and Wav2Vec2.0.

BERT utilizes the standard model from HuggingFace[2] [19], processing test taker responses through token embeddings. Each token is transformed into vector form and processed by BERT's encoder. The classification system includes three dense layers of 768 units each, followed by three layers of 128 units. The final output uses a 5-unit softmax layer for multi-class classification. Training involves 25 epochs using an AdamW optimizer, with a batch size of 8 and a learning rate of 5e-5, keeping the BERT layer frozen.

Wav2vec2.0 employs pre-trained Wav2vec2.0[3] [19] to convert audio signals into vector representations. Mean pooling aggregates vectors from the last hidden layer, adjusting for variable audio lengths. This feeds into a 768-unit dense layer, then an output layer. The model trains over 8 epochs

[2]https://huggingface.co/google-bert/bert-base-uncased
[3]https://huggingface.co/facebook/wav2vec2-base

Model	Test		Unseen Test	
	Accuracy(%)	Weighted-F1	Accuracy(%)	Weighted-F1
Qian2019 [15]-replicated	55.56	0.548	64.33	0.628
BERT	57.78	0.559	68.00	0.659
Wav2vec2.0	56.67	0.557	61.67	0.602
Multimodel [13]-replicated	64.44	0.639	66.67	0.650
SAMAD ($c = 1.2$)	**65.56**	**0.648**	**69.67**	0.684
w/o soft labels	63.33	0.636	**69.67**	**0.686**

Table 2. Comparison of baseline models on the GEPT datasets. The best results are marked in bold.

with an AdamW optimizer, a batch size of 4, gradient accumulation every 2 steps, and a learning rate of 1e-3, with the Wav2vec2.0 layer remaining frozen.

Qian2019 Model [15] bears some resemblance to SAMAD, which adopts three sub-modules: content, delivery, and language use. Each module uses Bi-LSTMs to embed the coming input features. The feed-forward attention layer is then applied to enable each sub-model to pay attention to the classification layer. Three subscores are concatenated and fed into a dense layer to predict a holistic score for each response.

Multi-modalities model integrates BERT and Wav2Vec2.0 [13], accepting dual inputs: transcriptions from ASR and WAV files. Each input is processed separately to generate representations, which are then combined via cross-modality attention. The unified representation is used to compute the proficiency score. The model trains using the AdamW optimizer for 8 epochs, with a batch size of 4 and gradient accumulation every 2 steps, maintaining a learning rate of 1e-5. During training, the layers of BERT and Wav2Vec2.0 are kept frozen to preserve the pre-trained parameters.

5.2. Implementation details

In our study, we utilized the pre-trained Whisper Medium model [20], known for its superior performance in multilingual ASR, even with unfamiliar datasets. Despite its strengths, Whisper provides only textual transcriptions without timestamps. To overcome this, we suggest using WhisperX [21], an enhancement designed to extract detailed word timestamps, thus enriching the data utility for more complex ASR applications. With WhisperX ASR, we obtained 0.38 word error rate (WER) on the LTTC GEPT Intermediate-level dataset. This result was based on a random sampling of 7% of the responses, which included 84 responses.

For the SAMAD proposed model, we use a pre-trained BERT to encode words in responses and prompts within the content sub-module. In the delivery and language use sub-modules, word representations are captured by 14-dimensional and 246-dimensional vectors, respectively. Each aspect's response-level representation is subsequently processed through a series of one-dimension convolution layers and a multi-head self-attention layer, consisting of 3 layers, with each layer having 4 heads and a hidden size of 256. Subsequently, these representations are fed into cross-aspect modules, which dynamically focus on features across different aspects. Finally, the outputs from the cross-aspect model are concatenated and passed through a self-attention transformer and multi-dense layer to make predictions. Training parameters were set with an AdamW optimizer, a batch size of 8, and a learning rate of 1e-4, while BERT's parameters were frozen. We evaluated the model's performance using classification accuracy and weighted F1 scores as metrics.

5.3. Results on Performance Comparison

As illustrated in Table 2, SAMAD surpasses four baseline models on both test and unseen data. The F1-measure evaluations reveal a consistent trend. These results imply that applying soft labels and combining SSL with handcrafted indicator features contribute to the improved performance of SAMAD.

In both the seen and unseen test datasets, it might initially be surprising that Qian2019 and BERT outperform the Wav2vec2.0 model. We hypothesize that this outcome can be attributed to the assessment criteria at the Intermediate level, which primarily focus on content and topic development features. Since the test takers are intermediate-level learners whose delivery is already sufficiently proficient, learners and human raters will more likely focus on language form and topic development rather than delivery features. This finding may supported by [22] that the higher the proficiency, the more processing room that learners can devote to language form. Therefore, the distinguishing capabilities of the Wav2vec2.0 model, which are more sensitive to nuances in delivery, become less impactful for intermediate-level learners as the focus shifts towards grammatical accuracy and topic coherence.

5.4. Effectiveness of Soft-label Optimization

We conducted an ablation study to evaluate the impact of soft-label optimization on SAMAD's performance. The model achieved optimal performance at $c = 1.2$. This result is derived from considering its performance on both the test and unseen test sets. This parameter was established as the benchmark for comparisons with baseline models. Soft-label optimization significantly boosted performance on both seen and

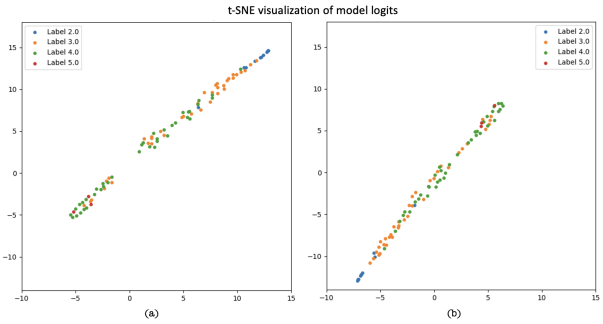

Fig. 3. Visualization of model logits using t-SNE on seen test data: (a) without soft-label optimization, and (b) with soft-label optimization.

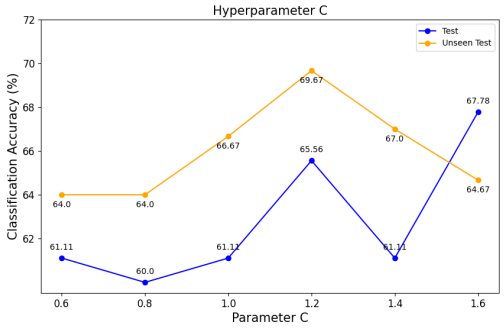

Fig. 4. Classification accuracy of SAMAD over different values of the hyperparameter c for soft-label optimization.

unseen test datasets, demonstrating that soft labels are essential in enhancing the efficacy of SAMAD.

In Figure 3, t-SNE visualizations demonstrate the impact of using soft-label optimization on the SAMAD model in seen test sets. Figure 3 (a), without soft-label optimization, shows clear boundaries that may not effectively handle overlapping categories. Conversely, Figure 3 (b) with soft-label optimization displays smoother transitions between classes, illustrating the model's improved ability to manage data ambiguity and adapt to real-world category scenarios. This highlights soft labels' role in enhancing model flexibility and aligning more closely with human rater judgments.

5.5. Data Distribution Influence

As demonstrated in Table 2, it is noteworthy that the model generally performs better on the unseen test set compared to the seen test set. This variation can be primarily attributed to the differences in data distribution across the dataset. Specifically, the distribution of the unseen test set more closely aligns with that of the training data, as depicted in Figure 5. This similarity in class distribution likely enhances the model's ability to generalize better on the unseen test set. Furthermore, when applying various configurations of soft-label optimization, we observe fluctuations in the results on the seen data; however, the performance on the seen test dataset consistently remains stable, as illustrated in Figure 4. Based on these observations, we can infer that data distributions are the primary reasons why the model tends to perform better on the unseen test dataset than on the seen test dataset. Despite the differences in data distribution, the SAMAD model demonstrates great performance improvements on both seen test and unseen test datasets, unaffected by the differences in data distribution. This robustness highlights the applicability and effectiveness of SAMAD in managing diverse data characteristics and maintaining consistent performance across varying spoken.

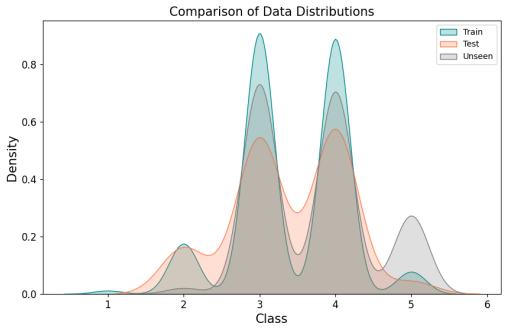

Fig. 5. Comparative visualization of class densities in train, seen test, and unseen test datasets.

6. CONCLUSION

In this paper, we have proposed a novel ASA modeling framework (dubbed SAMAD), which can harness the synergistic power of soft-label optimization and self-supervised learning (SSL) features in concert with handcrafted features. SAMAD demonstrates a marked improvement in effectiveness compared to baseline models. As a side note, this study acknowledges certain limitations, including the potential for further exploration of speech features and the integration of more advanced models. Future efforts will focus on developing a more comprehensive interpretability framework to assist learners or test-takers in understanding and improving their skills.

7. ACKNOWLEDGEMENTS

This work was supported by the Language Training and Testing Center, Taiwan. Any findings and implications in the paper do not necessarily reflect those of the sponsor.

8. REFERENCES

[1] Stefano Bannò and Marco Matassoni, "Proficiency assessment of l2 spoken english using wav2vec 2.0," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 1088–1095.

[2] Mao Saeki, Yoichi Matsuyama, Satoshi Kobashikawa, Tetsuji Ogawa, and Tetsunori Kobayashi, "Analysis of multimodal features for speaking proficiency scoring in an interview dialogue," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 629–635.

[3] Seongjin Park and Rutuja Ubale, "Multitask learning model with text and speech representation for fine-grained speech scoring," in *2023 IEEE Automatic Speech Recognition and Understanding, ASRU 2023 - Proceedings*, 2023.

[4] Jared Bernstein, Michael Cohen, Hy Murveit, Dimitry Rtischev, and Mitch Weintraub, "Automatic evaluation and training in english pronunciation," in *ICSLP*, 1990.

[5] Catia Cucchiarini, Helmer Strik, and Lou Boves, "Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech," *The Journal of the Acoustical Society of America*, vol. 111, pp. 2862–73, 07 2002.

[6] Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken english," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009, Spoken Language Technology for Education.

[7] Eesung Kim, Jae-Jin Jeon, Hyeji Seo, and Hoon Kim, "Automatic pronunciation assessment using self-supervised speech representation learning," 04 2022.

[8] Jidong Tao, Shabnam Ghaffarzadegan, Lei Chen, and Klaus Zechner, "Exploring deep learning architectures for automatically grading non-native spontaneous speech," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6140–6144.

[9] Ryuki Matsuura, Shungo Suzuki, Mao Saeki, Tetsuji Ogawa, and Yoichi Matsuyama, "Refinement of utterance fluency feature extraction and automated scoring of l2 oral fluency with dialogic features," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 1312–1320.

[10] Yao Qian, Rutuja Ubale, Matthew Mulholland, Keelan Evanini, and Xinhao Wang, "A prompt-aware neural network approach to content-based scoring of non-native spontaneous speech," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 979–986.

[11] Lei Chen, Jidong Tao, Shabnam Ghaffarzadegan, and Yao Qian, "End-to-end neural network based automated speech scoring," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6234–6238.

[12] Yaman Kumar Singla, Jui Shah, Changyou Chen, and Rajiv Ratn Shah, "What do audio transformers hear? probing their representations for language delivery structure," in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2022, pp. 910–925.

[13] Jiajun Liu, Aishan Wumaier, Cong Fan, and Shen Guo, "Automatic fluency assessment method for spontaneous speech without reference text," *Electronics*, vol. 12, no. 8, 2023.

[14] A. Green, C. Inoue, and F. Nakatsuhara, "Relating gept speaking tests to the cefr," LTTC–GEPT Research Report RG-09, LTTC, Taipei, 2017.

[15] Yao Qian, Patrick Lange, Keelan Evanini, Robert Pugh, Rutuja Ubale, Matthew Mulholland, and Xinhao Wang, "Neural approaches to automated speech scoring of monologue and dialogue responses," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 8112–8116.

[16] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019, pp. 6558–6569, Association for Computational Linguistics.

[17] Jinshan Zeng, Yudong Xie, Xianglong Yu, John Lee, and Ding-Xuan Zhou, "Enhancing automatic readability assessment with pre-training and soft labels for ordinal regression," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, Eds., Abu Dhabi, United Arab Emirates, Dec. 2022, pp. 4557–4568, Association for Computational Linguistics.

[18] Raul Diaz and Amit Marathe, "Soft labels for ordinal regression," 06 2019, pp. 4733–4742.

[19] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin

Lhoest, and Alexander Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Qun Liu and David Schlangen, Eds. Oct. 2020, pp. 38–45, Association for Computational Linguistics.

[20] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[21] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," 2023.

[22] Masatoshi Sato and Kim McDonough, "Predicting l2 learners' noticing of l2 errors: Proficiency, language analytical ability, and interaction mindset," *System*, vol. 93, pp. 102301, 2020.