Layer-Wise Feature Distillation with Unsupervised Multi-Aspect Optimization for Improved Automatic Speech Assessment

Chung-Wen Wu* and Berlin Chen†
* † Department of Computer Science and Information Engineering, National Taiwan Normal University
* E-mail: 40947040S@ntnu.edu.tw
† E-mail: berlin@ntnu.edu.tw

***Abstract*—Self-supervised features have shown promising progress across several domains. In Automatic Speech Assessment (ASA), SSL features have been widely utilized in recent research. However, few studies have dedicated efforts to explore the layer-wise features in pre-trained SSL models. Another key challenge in ASA is the high cost of labeling various aspects of speech proficiency, such as content relevance, delivery, and language use. In this paper, we propose three unsupervised subtasks to assist model training in ASA and examine the importance of embeddings from each layer of the acoustic model for various aspects. This provides preliminary research in this area. Extensive experiments demonstrate that model training with our tailored subtasks achieves superior performance in speech proficiency assessment tasks.**

I. Introduction

Automatic Speech Assessment (ASA) for second language (L2) proficiency offers significant benefits for both learners and educators. For L2 learners, ASA provides timely and feedback, which is crucial for effective language acquisition and skill improvement. This immediate response helps learners identify and correct errors, reinforcing proper pronunciation. For educators and testing institutions, ASA ensures more efficient and consistent scoring of language assessments. By automating the evaluation process, ASA reduces human error and bias, leading to fairer and more accurate results.

There are various methods to measure the speech proficiency of L2 learners, which can be broadly categorized into three components: content, delivery, and language use. Content assessment evaluates the relevance and coherence of the spoken material. Delivery encompasses prosody, pronunciation, stress, and other aspects related to how the speech is articulated. Language use focuses on grammar, morphology, and syntactic dependency, assessing the accuracy and complexity of the language structures employed.

However, labeling content, delivery, and language use separately is extremely time-consuming and requires professional training for the labelers. Moreover, recent research indicates that detailed scores assigned by experts may exhibit more bias compared to holistic scores.

Early studies in ASA [1]–[4] predominantly relied on handcrafted features related to different aspects of speech proficiency as input, ultimately making only holistic score predictions. While these features offer interpretability and align directly with human grading criteria, they are inherently limited by the holistic label. A holistic label alone is insufficient for providing comprehensive information necessary to effectively learn representations of various aspects of speech proficiency.

SSL representations have made significant strides across a range of speech processing tasks [5]–[10], including automatic speech recognition (ASR), speech enhancement, keyword spotting, speaker diarization, mispronunciation detection and diagnosis (MDD), and ASA. Models like wav2vec 2.0 [5], Whisper [11], and HuBERT [12], pre-trained on extensive datasets, leverage contextual information to extract features that are more robust and generalizable. In the realm of ASA, these models have proven highly effective in representing advanced speech-related features [13], thereby enhancing the accuracy of speech assessment.

In this paper, we propose a novel model architecture to explore SSL features across different layers of pre-trained acoustic encoders. We tailor various unsupervised subtasks for content, delivery, and language use to leverage the potential abilities of SSL features. By doing so, we provide more detailed information to enhance the training of models for ASA, avoiding the enormous consumption of human resources and ensuring a more comprehensive evaluation of L2 proficiency.

Extensive experiments demonstrate that model training with our tailored subtasks achieves superior performance in speech proficiency assessment tasks. Furthermore, our results show that information related to content, delivery, and language use is distributed across different layers of the acoustic encoder.

In summary, this paper presents three main contributions:

1) Tailoring unsupervised subtasks for content, delivery, and language use provides more concrete information than relying solely on holistic score labels, while also minimizing the need for additional human resources.
2) Investigating and analyzing the feature distribution of each aspect related to speech proficiency in the acoustic encoders.
3) Proposing a model trained with our tailored subtasks, we demonstrate that our approach significantly improves the accuracy of ASA tasks.

To the best of our knowledge, the relationship between speech proficiency features and latent representations at each layer of the acoustic encoder in ASA is an underexplored topic. This paper aims to provide preliminary research in this area.

II. Related Work

Recently, several studies in ASA have been conducted [14]–[16]. However, due to the scarcity of publicly available resources, these studies typically use only holistic labels as targets for model training. Although some research has access to detailed labels, such as topic development, delivery, and language use, this data is nonpublic. Even if we invest in and train human raters to label such detailed data, the collected data may not be sufficiently reliable due to the halo effect, where raters may rate all or some aspects of speech proficiency with similar scores.

SSL features from models like Whisper, wav2vec 2.0, and HuBERT have demonstrated great performance in several studies. However, to date, only a few studies have leveraged the features from the various layers of these pre-trained models to make assessments in ASA systems.

III. Methodology

A. Pseudo label generation

We generate pseudo labels for content, delivery, and language use from the data. These pseudo labels provide more detailed and stable information related to speech proficiency assessment, eliminating the need for human raters.

1) Content pseudo label: In content, we direct using content number associated with data as label.

2) Delivery pseudo label: In delivery, to measure a user's fluency and pronunciation, we begin by extracting some handcrafted features related to delivery and use the K-means algorithm to cluster the data into k clusters, which we then use as delivery pseudo labels. We employ two different types of acoustic models: monolingual wav2vec 2.0 and multilingual wav2vec 2.0. We start by using multilingual wav2vec 2.0 to recognize the transcriptions of user speech. Since L2 learners' pronunciation and accents are quite diverse, the multilingual training model can achieve better accuracy than the monolingual model in this scenario. On the other hand, we utilize monolingual wav2vec 2.0 to segment the audio signal into word-level segments using the CTC segmentation algorithm for force alignment with ASR transcriptions recognized by multilingual wav2vec 2.0. Using the monolingual model for alignment implies that if the L2 learner pronounces more accurately, the alignment result will be more correct. Within each word-level segment, we extract various delivery-related features, including pitch, duration (DUR), intensity, following silence, posterior probability from both monolingual and multilingual wav2vec 2.0, LM score from the n-gram LM during transcription, and confidence score from multilingual wav2vec 2.0. These features are utilized to construct an 8-dimensional continuous segment feature d_i, and subsequently, a sequence of segment features $[d_1, d_2, \ldots, d_M]$ is concatenated to represent the delivery feature, where M is a predefined length achieved by either truncating or padding with zeros. Finally, we use the K-means algorithm to obtain k clusters as pseudo labels.

3) Language use pseudo label: This part focuses on analyzing the grammar and syntax of sentences spoken by the user. We obtain the ASR transcription during the delivery pseudo label process. We extract language use features from each word to construct a sequence of language use features, which is then constrained to length M by padding or truncating. Next, we use spaCy, a natural language processing toolkit in Python, to extract Part-Of-Speech (POS) tags, dependency labels (DEP), and morphology (Morph.) labels. These features are encoded using one-hot encoding to obtain sparse language use features. Finally, similar to the delivery pseudo label generation, we leverage the K-means algorithm to cluster the data into k clusters as pseudo labels.

B. Proposed Model

To explore and leverage the representations of each layer in acoustic models, we propose a model architecture as illustrated in Fig. 1. This architecture trains three weighted vectors, w_c, w_d, and w_l, which separately combine acoustic embeddings into latent embeddings for content, delivery, and language use aspects. Subsequently, we use a linear layer to adapt the combined embeddings into latent representations with H dimensions separately. We employ CLS token embeddings to concatenate with combined latent representations, which are then passed through bidirectional Transformer blocks [17] of depth D separately. We extract the CLS token representation after the Transformer blocks to predict pseudo labels for content, delivery, and language use using multilayer perceptron (MLP) with H_o hidden dimensions to assist training. Finally, all CLS tokens are concatenated and using MLP to make the final holistic score assessment.

C. Training process

Based on our observations, although pseudo labels can provide rich information to assist holistic score prediction, they still contain some noise due to forced alignment errors or ASR transcript errors. To mitigate these side effects, we leverage a triplet loss approach across the three subtasks. This method helps learn well-represented continuous embeddings that cluster similar pseudo labels together and separate dissimilar ones. The benefit is a reduction in noise effects and the development of more robust representations. Finally, the holistic score is predicted using cross-entropy loss to train the model. The formulas are as follows:

$$L = L_c + \lambda(L_t(a_c, p_c, n_c) + L_t(a_d, p_d, n_d) + L_t(a_l, p_l, n_l)) \tag{1}$$

where L_c is cross entropy loss, L_t is triplet loss, a_c, p_c, n_c are anchor, positive and negative samples respectively.

IV. Dataset

Public corpora in ASA are quite scarce, with the most popular and commonly used datasets being ICNALE and Speechocean762. Both have their strengths and weaknesses.

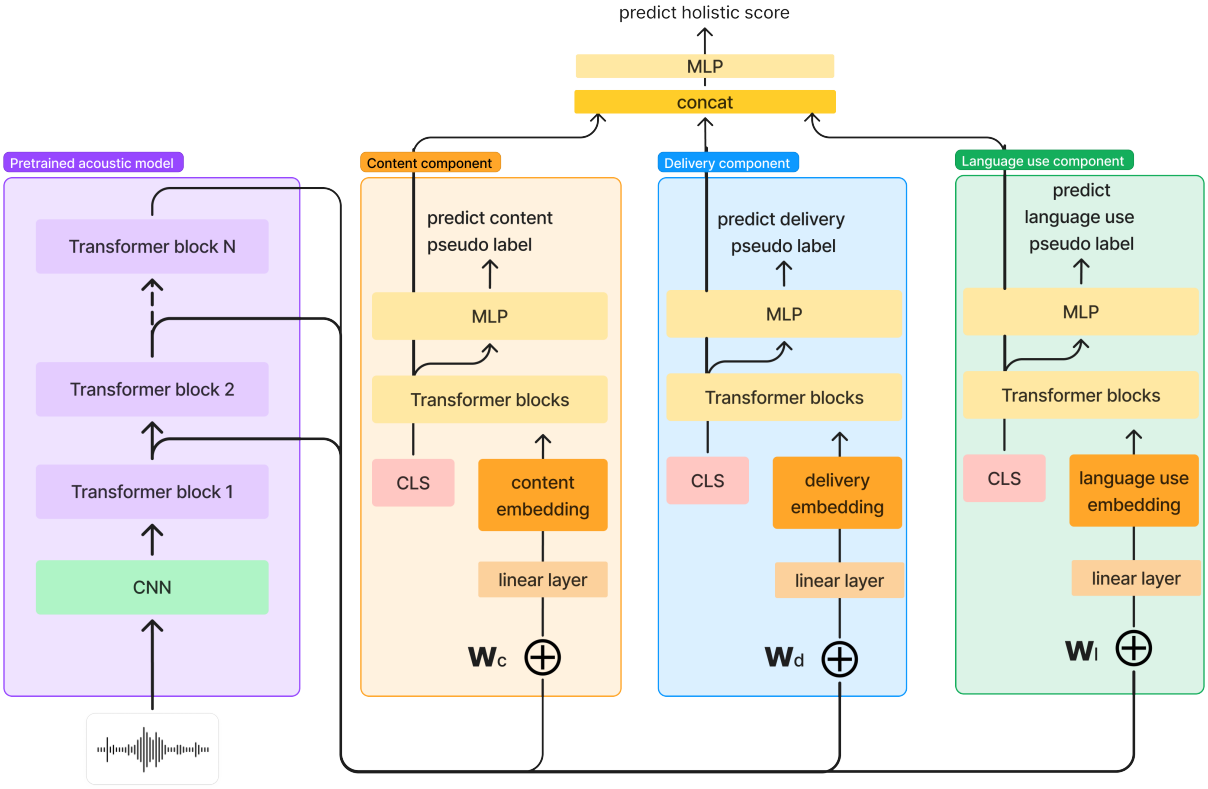

Fig. 1. The proposed model architecture combines latent representations from each layer of a pre-trained acoustic model to obtain embeddings for various aspects of speech proficiency. Leveraging pseudo labels to assist the training process.

The spoken part of the ICNALE dataset includes adequately lengthy monologues and dialogues, with each monologue being roughly 60 seconds and each dialogue lasting 30 to 40 minutes. However, ICNALE uses English test scores such as TOEFL, TOEIC, IELTS, and others as labels. These standard tests reflect the vocabulary and some reading skills of English learners but may not be sufficiently related to speaking skills.

In contrast, the Speechocean762 dataset has a large amount of data and detailed score labeling, such as accuracy, completeness, fluency, prosody, and even phoneme-level labeling. However, the speech in this dataset is quite short, typically around 2 to 5 seconds, and consists only of read-aloud tasks. This limitation prevents a comprehensive assessment in ASA systems.

Due to the above reasons, we used a corpus collected by the Language Training and Testing Center from the General English Proficiency Test (GEPT) intermediate level exam in this study. GEPT is an English certification test in Taiwan that covers listening, reading, writing, and speaking. For our study, we use data from the speaking section. This exam is a high-stake English proficiency test. It includes 1,199 responses divided equally among four sets of questions, each set answered by a different participant. The instructions given to participants were: "Below are a picture and four related questions. Please complete your answers in one and a half minutes. Do not read the number or the question when you answer. Please first look at the picture and think about the questions for thirty seconds."

Figure 3 illustrates the distribution of scores and response lengths. The scores, which range from 1 to 5, were calculated as averages from assessments provided by two professionals, with any floating-point values being discarded unconditionally. To create an unknown content test set, we randomly selected a specific set of questions. This set was designed to assess performance under cold start conditions, where the model encounters previously unseen content. The data not used for the unknown content test set was split into training, development, and known content test sets in an 8:1:1 ratio. This approach guarantees that the model receives substantial training data while also enabling thorough evaluation on both familiar and new content types.

V. Experiments and Results

A. Experiments Setup

The pretrained acoustic model used in our experiments is the Whisper-base model, which has 7 Transformer blocks in the encoder. The depth of the Transformer blocks in our model is set to 3. The dimensions for the latent representations in each aspect (H) and the hidden layers in the MLP (H_o) were configured to be 128 and 64, respectively. Throughout the training process, we employed the RAdam optimizer, incorporating a weight decay of 1×10^{-5}. The learning rate

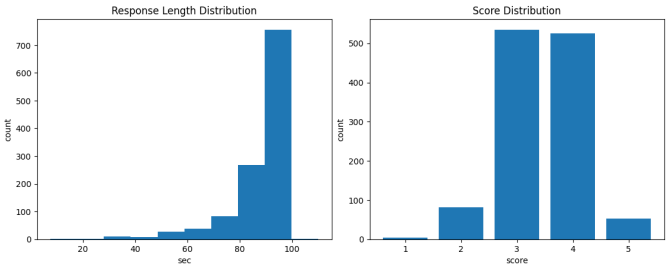

Fig. 2. Response length and score distribution in the GEPT intermediate level dataset.

was set to 2×10^{-4}, and all experiments were carried out over 16 epochs with a batch size of 8.

B. Ablation studies

TABLE I
ABLATION STUDIES ON THREE COMPONENTS

Model	Known Content Accuracy	Unknown Content Accuracy
Complete model	**0.722**	**0.727**
Without language use	0.667	0.697
Without delivery	0.644	0.697
Without content	0.667	0.683
Without any subtask[a]	0.633	0.647

[a] Without any subtask means that we only use a combined embedding from the acoustic model. We then sequentially pass this combined embedding through the same blocks of the model as illustrated in Fig. 1, extracting a CLS token embedding to predict the holistic score.holistic score predict.

As shown in Table I, the experiments demonstrate that training with three subtasks outperforms all other ablation models, indicating that each subtask improves the model's accuracy in holistic score assessment. The experiments also highlight that directly using the holistic score for training or combining embeddings from each layer for holistic score prediction is not optimal.

C. Visualize the weight distributions in combined embeddings

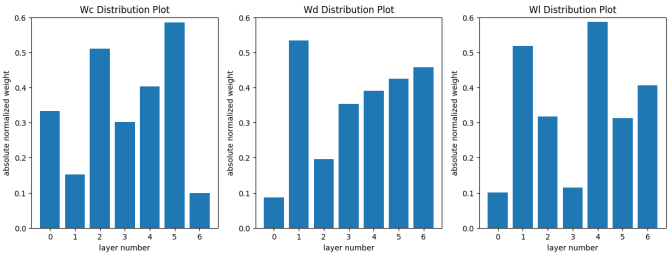

Fig. 3. Absolute weight distributions across each layer in the whisper-base encoder

We extract w_c, w_d, w_l from complete model, normaling it and take absolute value to shows each layers importance in whisper-base encoder, as shown in Fig. 3. Overall, the distributions for each aspect are quite distinct, proving that latent

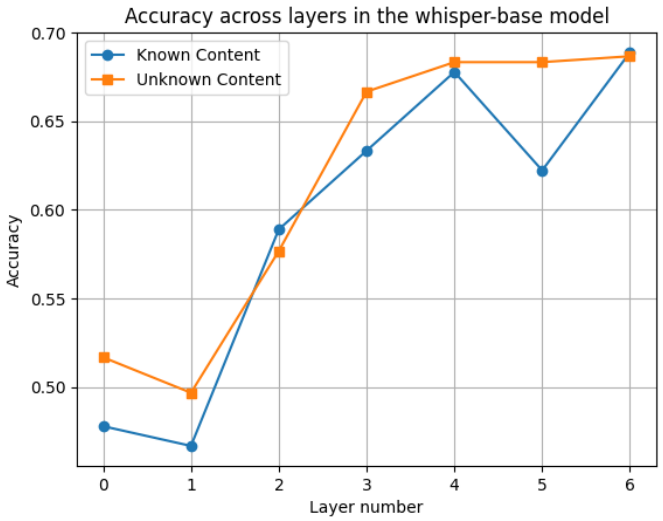

Fig. 4. We directly extract embeddings from each layer and sequentially pass them through the same blocks of the model, as illustrated in Fig. 1, to make holistic score predictions. The layer numbers range from 0 to 6, representing layers in the encoder from shallow to deep.

representations in different layers contain different meanings. Our method successfully directs attention to these various aspects, leveraging the unique information captured at each layer. This differentiation enhances the model's ability to make more accurate and nuanced assessments. For the content aspect, the features tend to focus on the middle layers. In the delivery aspect, layer 1 is the most significant, but other important features tend to focus on the deeper layers. For the language component, the features tend to focus on both shallow and deep layers.

D. Performance using features from each layer directly

In the Whisper-base model, when training using only the holistic score, the general tendency is that deeper layers provide better performance, as shown in Fig. 4. Experiments indicate that training with only the holistic score tends to yield better performance when using the last layer alone, achieving 0.689 and 0.687 in known content and unknown content, respectively. In contrast, combining all layers in the Whisper-base model, as shown in Table I, results in scores of 0.633 and 0.647 in known content and unknown content, respectively. This phenomenon points out that the holistic score alone may not provide sufficient information to effectively train a well-represented combined latent representation.

VI. CONCLUSIONS

We proposed a training method with three subtasks and demonstrated that this method can benefit model training in the ASA system using SSL features. By incorporating subtasks for content, delivery, and language use, our approach leverages rich pseudo labels to provide detailed and consistent information. This method improves the robustness and accuracy of the model, ensuring a more comprehensive assessment of L2 proficiency. The effectiveness of our approach is validated

through extensive experiments, demonstrating superior performance compared to relying solely on holistic labels.

Our experiments show that features of various aspects emerge in different layers of the model. This observation highlights the importance of leveraging multi-layer embeddings to capture the diverse and complex characteristics of speech proficiency. By analyzing and utilizing the representations from multiple layers, our proposed method can more effectively leverages content, delivery, and language use, leading to improved performance in ASA systems.

References

[1] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[2] Y. Qian, P. Lange, K. Evanini, *et al.*, "Neural approaches to automated speech scoring of monologue and dialogue responses," in *Proceedings of the 2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2019, pp. 8112–8116.

[3] Z. Yu, V. Ramanarayanan, D. Suendermann-Oeft, *et al.*, "Using bidirectional lstm recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech," in *Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 338–345. DOI: 10.1109/ASRU.2015.7404814.

[4] L. Chen, J. Tao, S. Ghaffarzadegan, and Y. Qian, "End-to-end neural network based automated speech scoring," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6234–6238. DOI: 10.1109/ICASSP.2018.8462562.

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[6] W.-N. Hsu, A. Sriram, A. Baevski, *et al.*, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *arXiv preprint arXiv:2104.01027*, 2021.

[7] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[8] H. S. Bovbjerg and Z.-H. Tan, "Improving label-deficient keyword spotting through self-supervised pre-training," in *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSP)*, 2023, pp. 1–5.

[9] X. Xu, Y. Kang, S. Cao, B. Lin, and L. Ma, "Explore wav2vec 2.0 for mispronunciation detection.," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2021, pp. 4428–4432.

[10] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, "Multimodal emotion recognition with transformer-based self supervised feature fusion," *IEEE Access*, vol. 8, pp. 176 274–176 285, 2020. DOI: 10.1109/ACCESS.2020.3026823.

[11] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of International Conference on Machine Learning*, PMLR, 2023, pp. 28 492–28 518.

[12] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[13] S. Park and R. Ubale, "Multitask learning model with text and speech representation for fine-grained speech scoring," in *Proceedings of the 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–7.

[14] E. Kim, J.-J. Jeon, H. Seo, and H. Kim, "Automatic pronunciation assessment using self-supervised speech representation learning," *arXiv preprint arXiv:2204.03863*, 2022.

[15] J. Park and S. Choi, "Addressing cold start problem for end-to-end automatic speech scoring," *arXiv preprint arXiv:2306.14310*, 2023.

[16] S. Bannò and M. Matassoni, "Proficiency assessment of L2 spoken english using wav2vec 2.0," in *Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 1088–1095.

[17] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.